Development of a working memory test for the German Bundeswehr's online assessment

D I S S E R T A T I O N
zur Erlangung des akademischen Grades

Doctor rerum naturalium
(Dr. rer. nat.)

eingereicht an der
Lebenswissenschaftlichen Fakultät der Humboldt-Universität zu Berlin

von
M.Sc. Ursa Nagler-Nitzschner, geb. Nagler

Präsidentin
der Humboldt-Universität zu Berlin

Prof. Dr. Ing. Dr. Sabine Kunst

Dekan der Lebenswissenschaftlichen Fakultät
der Humboldt-Universität zu Berlin

Prof. Dr. Bernhard Grimm

Gutachter
1. Prof. Dr. Matthias Ziegler
2. Prof. Dr. Martin Brunner

Tag der mündlichen Prüfung: 22.02.2021

“One of the most successful undertakings attributed to modern psychology is the measurement of mental abilities.”

Kevin Lamb (1994, p. 386)

Acknowledgments

First of all, I want to thank Professor Dr. Matthias Ziegler. I was overjoyed when you accepted me as your PhD student and was always grateful for your valuable input and advice. I also want to thank Professor Dr. Martin Brunner for taking the time to serve as my second supervisor.

Although he would probably deny that he deserves to be mentioned here, I want to thank Professor Dr. John Rauthmann. John, you had a huge influence on me in my first years of university and have always served as a supportive role model.

I also want to thank Professor Dr. Sidney Irvine for consultation during the early stages of the project. Special thanks to LRDir a. D. Johannes Wulf and to my former supervisor, RDir'in Dr. Sibylle Dunker, for committing herself to this project. Furthermore, I owe my gratitude to the whole team at BMVg P III 5 for helping me navigate the bureaucratic obstacles. LRDir Bernd Völker, RDir Florentin Klein and RDir'in Dr. Andrea Heiß deserve particular mention. I am indebted as well to all staff members at the Bundeswehr Career Centre who helped me conduct the study, to my colleagues Sylvia Weber and Marko Vidaković, and to the participants for voluntarily participating. My thanks also go to the Assessment Centre for Bundeswehr Officers for the opportunity to test during ongoing operations.

Of course, I also want to thank my family, especially my parents. Your support over the years has been invaluable – I hope I have made you proud.

Last but not least, a huge thank you goes to my understanding husband and daughter. You encourage me to set and achieve high goals even when I doubt myself.

Zusammenfassung

Es ist seit mehreren Jahren bekannt, dass militärische Organisationen mit Nachwuchsproblemen zu kämpfen haben (z.B. Harris, 2018/2018; Koker, 2019/2019; Squires, 2019/2019; The Local, 2019/2019; Wolfgang, 2019). Die Bundeswehr stellt hierbei keine Ausnahme dar (Handelsblatt, 2019; Jungholt, 2018/2018). Dabei ist das Nachwuchsproblem kein militärspezifisches. Generell ist der Arbeitsmarkt im Begriff, sich weg von einer hohen Nachfrage seitens der Arbeitnehmerinnen und Arbeitnehmern nach Arbeitsplätzen hin zu einem Offerieren von Arbeitsangeboten seitens des Arbeitgebers zu entwickeln. Dies lässt sich durch die sinkende Anzahl an Fachkräften bei gleichzeitig steigendem Bedarf nach diesen erklären. Aus diesem Grund steigt der Konkurrenzdruck in der Anwerbung von Nachwuchskräften bei Unternehmen. Dieses Phänomen ist auch unter dem Begriff "war for talents" (Busold, 2019) bekannt. Aus diesem Grund ist es von höchstem unternehmerischem Interesse, kompetentes Personal zum frühestmöglichen Zeitpunkt zu werben und an die eigene Organisation zu binden. Gleichwohl ist die Ausbildung von Personal mit hohen Kosten verbunden und falsche Personalentscheidungen können deshalb langwierige Konsequenzen nach sich ziehen. Damit wird eine effiziente als auch effektive Personalauswahl nötig. Zum einen, um unnötige Kosten zu vermeiden, was ebenfalls im höchsten Interesse der Anwenderinnen und Anwender ist (König, Klehe, Berchtold, & Kleinmann, 2010), und zum anderen, um den Auswahlprozess so kurz wie möglich zu gestalten, um so das Risiko zu minimieren, die Bewerberinnen und Bewerber in dieser Zeit an ein anderes Unternehmen zu verlieren. Für dieses Vorhaben sind durch das Internet neue Perspektiven und Ansätze geöffnet worden. Online Assessment oder e-Assessment (OA) nimmt bereits seit längerer Zeit einen bedeutenden Platz im Instrumentarium der Personalauswahl ein (z.B. Wiedmann, 2009) und scheint zukunftsfähig zu sein (Steiner, 2017). Jedoch existiert für die Personalauswahl der Bundeswehr derzeit noch kein OA.

Dieses könnte jedoch beispielsweise dafür genutzt werden, um den Bewerbungsprozess zu beschleunigen, in dem die aussichtsreichsten Bewerberinnen und Bewerber bevorzugt zur Präsenzdiagnostik eingeladen werden. Hierdurch wären sowohl ein schnellerer Auswahlprozess, als auch eine zügigere Bindung an das Unternehmen möglich.

Obwohl OA sehr viele Vorteile aufweist, sind auch einige Schwierigkeiten zu berücksichtigen. So wird das OA in einer Testumgebung mit fehlender Supervision durch Aufsichtspersonal durchgeführt, was es einfacher macht, bei den Testverfahren zu betrügen (Steger, Schroeders, & Gnambs, 2018) . Zur Entgegnung dieser Problematik existieren bereits vielfältige Ansätze: Diese reichen von adaptiven Testverfahren bis hin zur Implementierung großer Itempools, aus denen randomisiert Itemsets generiert werden. Allerdings geht die Erstellung von Items mit hoher psychometrischer Qualität auch mit hohen Kosten einher und ist gemessen an dem Bedarf an Items, den die Bundeswehr auf Grund einer sehr hohen Anzahl an Bewerberinnen und Bewerber hat (durchschnittlich 120 000 im Jahr, Handelsblatt, 2019), ineffizient. Der Ansatz der automatischen Itemgenerierung hingegen produziert Items auf Basis von Regeln, die a priori hinsichtlich ihrer Schwierigkeit evaluiert wurden. So können sehr viele Items kostengünstig und zeiteffizient erstellt werden. Hierfür wird ein passendes, latentes Konstrukt benötigt, das sich für automatische Itemgenerierung eignet und sinnvoll im Kontext der Personalauswahl ist. Da sich Intelligenz als der beste singuläre Prädiktor für Berufserfolg herausgestellt hat (z.B. Ree, Earles, & Teachout, 1994; Schmidt & Hunter, 1981, 1998; Schmidt, Oh, & Shaffer, 2016; Ziegler, Dietl, Danay, Vogel, & Bühner, 2011) und Arbeitsgedächtnis wiederum ein guter Prädiktor für Intelligenz ist (z.B. Gignac, 2014; Kane, Hambrick, & Conway, 2005; Oberauer, Schulze, Wilhelm, & Süß, 2005), scheint sich dieses Konstrukt in Bezug auf die Personalauswahl als vorteilhaft herauszustellen. Zudem eignet sich Arbeitsgedächtnis in hohem Maße für automatische Itemgenerierung, da die meisten Testverfahren, die Arbeitsgedächtnis messen,

repetitive Aufgaben beinhalten, die sich beispielsweise nur durch ihre Länge oder ihren semantischen Inhalt unterscheiden.

Ziel des vorliegenden Projektes war es deswegen, einen Arbeitsgedächtnistest mit einer hohen prädiktiven Validität für die Personalauswahl in der Präsenzdiagnostik für das OA der Bundeswehr zu entwickeln.

Die vorliegende Dissertation ist wie folgt aufgebaut: Zu Beginn wird die Personalauswahl in der Bundeswehr dargestellt. Diese Übersicht beinhaltet auch die Geschichte der Personalauswahl im militärischen Kontext sowie die Darstellung ausgewählte Testverfahren der Bundeswehr. Anschließend wird Überblick über OA gegeben und auf die mit OA assoziierten Risiken sowie entsprechende Lösungsansätze eingegangen. Des Weiteren werden das theoretische Fundament hinsichtlich des Arbeitsgedächtnisses aufgezeigt und hierfür relevante Modelle vorgestellt. Dies inkludiert auch die Arbeitsgedächtniskapazität. Dieses Kapitel schließt mit einer Vorstellung von Testverfahren, die zur Erhebung von Arbeitsgedächtnis eingesetzt werden. Im folgenden Kapitel werden dann die relevanten Grundlagen der Testtheorie erklärt, wobei automatische Itemgenerierung und Rasch Modelle spezielle Berücksichtigung finden.

Das nächste Kapitel beschäftigt sich mit der Entwicklung und Konstruktion der Testverfahren. Es wurden zwei Testverfahren entwickelt, die sich im Aufbau ähneln, sich jedoch in ihrem Fokus unterscheiden: während der erste Test den Fokus auf figurale Inhalte legt, liegt dieser beim zweiten Test auf figuralen-verbalen Inhalten. Den Abschluss des Kapitels bildet die dezidierte Ausführung der Ziele der vorliegenden Arbeit.

Im nächsten Abschnitt erfolgt dann die Darstellung von Studie 1 Das Ziel dieser (N = 330) war es zu zeigen, dass automatische Itemgenerierung für die vorgestellten und entwickelten Testverfahren überhaupt möglich ist. Zu diesem Zweck wurden fixe Itemsets in

einem Balanced Incomplete Block Design (Frey, Hartig, & Rupp, 2009) getestet. Die Itemsets wurden anschließend mit zwei variierenden Repräsentationszeiten der Stimuli getestet. Die Probandinnen und Probanden bearbeiteten schließlich ein Itemset. Um einen Vergleichsmaßstab zur Höhe des Zusammenhangs herzustellen, wurden zusätzlich mögliche Interkorrelationen der Parameter über eine Monte-Carlo-Simulation determiniert. Die Resultate weisen darauf hin, dass die automatische Itemgenerierung funktioniert, da die Betaparameter für dieses, eruiert durch ein Linear Logistisches Test Modell (LLTM), mit denen eines herkömmlichen Raschmodells hoch korrelieren. Gleiches gilt auch, wenn ein anderes Scoring angenommen wird und die Parameter von LPCM und PCM verglichen werden. Dabei schnitt die längere Repräsentationszeit (3 s) etwas besser ab, als die kürzere (1 s). Es zeichnete sich ab, dass das LLTM die Daten besser abbildet. Insgesamt konnte Studie 1 zeigen, dass die automatische Itemgenerierung für die entwickelten Tests möglich ist. Aus den Ergebnissen wurden Implikationen und Konsequenzen für die zweite Studie gezogen.

Im folgenden Kapitel wird anschließend die zweite Studie vorgestellt. Im Fokus dieser (N = 621) standen die Interpretation der Testwerte hinsichtlich konvergenter, divergenter und prädiktiver Validität, wie die interne Konsistenz und die Skalierung. Zu diesem Zweck absolvierten die Probandinnen und Probanden einen der zwei entwickelten Testverfahren sowie eine Digit Span backward. Zudem bearbeiteten sie die für die Personalauswahl nötigen Testverfahren. Auch hier zeigten sich wie in Studie 1 die besten Resultate für das LLTM, mit keinen nennenswerten Unterschieden zwischen beiden Testverfahren. Das beste Ergebnis zur Prädiktion der Präsenzdiagnostik konnte für die Offiziersbewerberinnen und –bewerber mittels des figuralen Testverfahrens erzielt werden.

Abschließend werden die Ergebnisse aus beiden Studien diskutiert und ein Fazit gezogen.

Summary

Nowadays, recruiting and therefore personnel assessment remain one of the military's primary concerns (e.g., Harris, 2018/2018; Koker, 2019/2019; Squires, 2019/2019; The Local, 2019/2019; Wolfgang, 2019), with the German military being no exception (Handelsblatt, 2019; Jungholt, 2018/2018).

The recruitment problem is not a specific problem of the military. In general, the labor market is changing from high demand for jobs on the part of applicants to high demand for applicants on the part of employers. This can be explained by the declining number of skilled workers and the simultaneous increase in demand for them. For this reason, companies are under increasing competitive pressure when it comes to recruiting. This phenomenon is known as the "war for talents" (Busold, 2019).

For this reason, it is of utmost corporate interest to recruit competent personnel at the earliest possible stage and to retain them in one's own organization. However, as the training of personnel is associated with high costs, incorrect personnel decisions can therefore have long-lasting consequences. Therefore, a efficient and effective personnel selection process is needed in order to avoid unnecessary costs, which is also in the highest interest of the organization (König et al., 2010), and make the selection process proceed as quickly as possible in order to minimize the risk of losing applicants to another company. The Internet has opened up new perspectives for this project. Online assessment or e-assessment (OA) has occupied an important place in the personnel selection toolbox for some time now (e.g., Wiedmann, 2009) and seems to be a solid option for future personnel selection purposes (Steiner, 2017). However, the Bundeswehr has yet to implement OA. This could be used, for example, to speed up the application process by giving preference to the most promising

applicants and inviting them to the on-site assessment. This would facilitate a faster selection process and thus a quicker commitment to the organization.

Although OA has a great number of advantages, there are also some difficulties to consider. For example, OA is conducted in an unsupervised testing environment without a live administrator, making it easier to cheat on tests (Steger et al., 2018). A variety of approaches already exist to counter this problem: ranging from adaptive test procedures to randmoly drawing item sets from extremely large item pools. However, creating items with a high psychometric quality is high-cost and therefore not feasible for the Bundeswehr due to the very large number of applicants (120 000 per year on average, Handelsblatt, 2019) and thus the need for a very large item pool.

The automatic item generation approach, on the other hand, produces items based on rules that have been evaluated a priori in terms of their difficulty. In this way, a large number of items can be generated in a cost-effective and time-efficient manner. This requires a latent construct that is suitable for automatic item generation and useful in the context of personnel selection. Since intelligence is the best singular predictor of job performance (e.g., Ree et al., 1994; Schmidt et al., 2016; Schmidt & Hunter, 1981, 1998; Ziegler et al., 2011) and working memory is a good predictor of intelligence (e.g., Gignac, 2014; Kane et al., 2005; Oberauer et al., 2005), working memory seems to be excellent for personnel selection purposes, especially for OA. Moreover, working memory is highly suitable for automatic item generation because most test procedures that measure working memory involve repetitive tasks that differ only in their length or semantic content, for example.

Therefore, the aim of the present project was to develop a working memory test for the Bundeswehr's OA with a high predictive validity for the outcome of the personnel selection process.

This dissertation is structured as follows: First, personnel selection in the Bundeswehr is presented. This overview also includes the history of personnel selection in the military context and the presentation of selected test procedures currently used by the Bundeswehr. Subsequently, an overview of OA is given and the risks associated with OA as well as corresponding solution approaches are discussed. Furthermore, theoretical foundations concerning working memory are introduced and relevant models are presented.

This includes the model of working memory capacity. The chapter concludes with a discussion of tests currently used to assess working memory. In the following chapter, the relevant basics of test theory are explained, with special attention to automatic item generation and Rasch models.

The next chapter deals with the development and construction of the tests. Two tests were developed that are similar in structure but differ in their focus: while the first test focuses on figural content, the second test focuses on figural-verbal content. The chapter concludes with a detailed description of the goals of the present study.

The next chapter then presents Study 1. The goal of this study with $N = 330$ was to show that automatic item generation works for both tests. For this purpose, multiple item sets were tested in a Balanced Incomplete Block Design (Frey et al., 2009).

The item sets were then tested with two different presentation times for the stimuli. The subjects completed one item set for each test. In order to establish a comparative measure of the correlational level, possible intercorrelations between the parameters were additionally determined via a Monte Carlo simulation study.

The results indicate that automatic item generation works, as the beta parameters for the Linear Logistic Test Model (LLTM) were highly correlated with those for a conventional

Rasch model. The same is true when a different scoring method is used and the parameters of a Linear Partial Credit Model (LPCM) and a Partial Credit Model (PCM) are compared.

The longer presentation time (3 s) performed slightly better than the shorter one (1 s). The LLTM represented the data better than the LPCM. Overall, Study 1 demonstrated that automatic item generation was feasible for the developed tests. Implications and consequences for the second study were drawn from the results.

The following chapter then presents the second study. The focus of the second study with $N = 621$ was to evaluate both tests for convergent, divergent and predictive validity as well as reliability and scaling.

For this purpose, the subjects completed one of the two developed tests as well as a Digit Span backward task. In addition, they completed the tests currently necessary for the Bundeswehr's personnel selection. Again, as in Study 1, the best results were for the LLTM, with no significant differences between the two test procedures. The figural test procedure was best able to predict officer applicant's on-site assessment result.

Finally, the results of both studies are discussed and a conclusion is drawn.

Contents

List of figures

List of tables

List of abbreviations

AC	Assessment Center for Bundeswehr Officers
CC	Career Center of the Bundeswehr
CHC	Cattell-Horn-Carroll model of cognitive abilities
CTT	Classical test theory
DIF	Differential item functioning
DS	Digit span
IFA	Item factor analysis
IRT	Item response theory
ISI	Interstimulus interval
LTM	Long-term memory
LLTM	Linear logistic test model
LPCM	Linear partial credit model
LRT	Andersen likelihood ratio test
OA	Online assessment / eAssessment
PCM	Partial credit model
RM	Rasch model
UIT	Unproctored Internet testing
WM	Working memory
WMC	Working memory capacity
WM-F	Figural working memory test
WM-V	Verbal working memory test

Introduction

Personnel assessment can look back on a long history. Testing of mental abilities goes back over 2,000 years to ancient China and has become more professional over time (Bowman, 1989). As early as during the Ming dynasty (1368-1644), formalized institutions for such evaluations existed (Bowman, 1989).

The history of personnel assessment in the military may be more recent, but is also tightly interwoven. As early as 1814, surgeons in the US Army were subjected to tests (DuBois, 1970). Even Francis Galton himself supported the British Royal Military Academy by applying his statistical concepts to admissions scores in 1869 (Stigler, 1999). The next major step was taken during World War I, as mass intelligence testing was developed in the US due to the need for new recruits, resulting in the Army Alpha Test (Embretson, 1999; Yerkes, 1921). However, psychological testing was popular not only in the US military, but also in the German (Salgado, Anderson, & Hülsheger, 2010; Sprung & Sprung, 2001), French (Salgado et al., 2010), British (Hearnshaw, 1964) and Italian (Salgado, 2001) militaries. With the testing efforts in the US Army declared a huge success, there was a testing boom in the US private sector after the war (Katzell & Austin, 1992). Nevertheless, psychological assessment in the military remained substantial, with psychology (and psychological assessment) becoming an integral part of Germany's military at this time (Fitts, 1946; Vinchur & Koppes Bryan, 2012). In the late 1930s, as World War II was approaching, assessment in the military peaked again (Ansbacher, 1941; Vinchur & Koppes Bryan, 2012). In the US, for example, the Army Alpha Test was replaced by the Army General Classification Test in 1940 (Harrell, 1992). During the 1940s, a progressive matrices test was used for military selection in the British military (Salgado et al., 2010), probably the first use of this test in the personnel assessment context. Psychological assessment remained essential for German military at that time as

well, with over 200 psychologists working for the military, mainly for selection purposes (Ansbacher, 1941). Many years later, in the 1980s, one of the largest and most expensive studies ever took place in the US military over a 7-year period (Borman, Klimoski, & Ilgen, 2003; Campbell, 1990), once again underscoring the close connection between assessment and the military. Another large-scale study was conducted by Lindqvist and Vestman (2011), who used data from Swedish military enlistees to test the predictive power of cognitive and non-cognitive tests for labor market outcomes like earnings or unemployment.

Nowadays, recruiting and therefore personnel assessment remain one of the military's primary concerns (e.g., Harris, 2018/2018; Koker, 2019/2019; Squires, 2019/2019; The Local, 2019/2019; Wolfgang, 2019), with the German military being no exception (Handelsblatt, 2019). Most western militaries are in competition with private-sector firms who can often offer more attractive jobs without the stresses and risks associated with being a soldier. With the job market shifting from excess of applicants to demand, high-potential personnel is particularly urgently needed and wanted (e.g., Busold, 2019), making the recruitment of high-potential employees difficult. The Internet and the ability to access it at any time through devices like tablets and smartphones necessitate a recruiting approach that is equally fast. Although there is little research on application withdrawal (Acikgoz & Sumer, 2019), an older study revealed that "time lags" are a major cause for withdrawal (Arvey, Gordon, & Massengill, 1975). Hence, an efficient application process is necessary. However, how can the correct person be chosen quickly? Classical personnel selection seems to provide an answer, but often takes a long time. In today's world, a faster approach is needed, and online assessment (OA) seems promising. Therefore, the goal of the present project was to develop a test fit for the German military's OA.

The follow section first describes the Bundeswehr's (German military) recruiting process in order to provide an overview of what to expect in the context of personnel selection in the German military. Second, a brief overview of OA in general is given, followed by an overview of working memory (WM) and test theory. Finally, automatic item generation is outlined.

Bundeswehr Recruiting and Personnel Selection

Nowadays, the Bundeswehr's posters and advertisements seem to be everywhere in Germany. The Bundeswehr seeks to maintain an ongoing media presence in the country, with its recruitment slogan "Mach, was wirklich zählt" (Do what really counts) visible at bus stops, at the mall or on the street. These advertisements promote a website (www.machwaswirklichzählt.de / www.bundeswehrkarriere.de) where interested persons can gain an overview of the different careers available in the military and make an appointment for an individual advisory session. At this session, the interested candidate is provided with all relevant documents to fill out in order to apply. An assessment date is set after the complete application is submitted (Bundeswehr, 2019b). It is not unusual to wait two to three months for an assessment appointment (Bundesamt für das Personalmanagement der Bundeswehr, personal communication, November 6, 2019). At this point, different personnel selection procedures take place depending on an applicant's selected career. For better understanding, the procedure for officer applicants is outlined here for illustrative purposes, because it is the longest and complex process.

Once their application has been processed, officer applicants are invited to a two-day assessment. Applicants arrive one day prior and receive all relevant information about the assessment in a presentation and fill out a demographic survey (Bundeswehr, 2014).

The first day entails an essay, medical examination, various psychological tests, and classic assessment center tasks like teamwork situations. The second day consists of fitness exams, sometimes further psychological tests, and advising on a course of studies[1] (Bundeswehr, 2012). As can be seen, the recruiting process takes quite a long time and involves high costs, as all applicants are reimbursed for travel costs and overnight accommodations are provided.

Psychological tests. All psychological psychometric tests involved in the German military's officer selection process are presented via computer (e.g., Oettershagen, 2015; Wagner & Klein, 2015) and are partially adaptive (Krex, 2008; Steyer & Partchev, 2000). It is common knowledge that intelligence is the best predictor of job performance when only one predictor is considered (Ganzach & Pankaj, 2018; Ree et al., 1994; Scherbaum, Goldstein, Yusko, Ryan, & Hanges, 2012; Schmidt et al., 2016; Schmidt & Hunter, 1981, 1998; Ziegler et al., 2011). This is unsurprising considering that different cognitive abilities correlate more strongly with each other as time goes on and are closely interwoven with one another (Breit, Brunner, & Preckel, 2020). In addition, intelligence is a predictor for key life outcomes in adulthood such as income (Hasl, Kretschmann, Richter, Voelkle, & Brunner, 2019) or health (e.g., Wrulich et al., 2014).

Therefore, it is in line with expectations that the Bundeswehr tests intelligence in its assessment process (Bundeswehr, 2019a). The Cattell-Horn-Carrol (CHC) model of cognitive abilities provides a good framework for understanding intelligence (Schneider & McGrew, 2012, 2018). Schneider and McGrew (2018) describe this model as follows: "It does not explain everything about intelligence, but it wants to" (p. 73). It understands

[1] All officers must complete a university degree in the German military before they begin working in their military occupational specialty.

intelligence as a range of different cognitive abilities which are grouped hierarchically and functionally (Schneider & McGrew, 2018). Therefore, different subtests within intelligence tests can be matched to different abilities within the CHC model. A detailed overview of the model is given in the section “Working memory models”.

To maintain the security of the testing material, it is not possible to outline every test within the Bundeswehr personnel selection process. However, the assessment training software provides a good overview (Bundeswehr, 2019a).

Nonetheless, three particular tests should be mentioned that, as presented below, are relevant for the present project: a verbal analogies test, an arithmetic test and a matrices test (Bundeswehr, 2016a, 2019a; Krex, 2008). These types of tests are frequently used to measure intelligence (Carpenter, Just, & Shell, 1990; Lynn, Chen, & Chen, 2011; Raven, 1981; Raven, Court, & Raven, 2008; Raven, Raven, & Court, 2003; Unsworth, 2010; Wechsler, 2008; Whitely, 1976).

As analogies are integral to human intelligence (Spearman, 1923, 1927; Sternberg, 1977), a verbal analogies test is administered. Since reading skills are closely linked to comprehension knowledge within the CHC model (Evans, Floyd, McGrew, & Leforgee, 2002), performance in verbal analogies probably reflects cognitive abilities like comprehension knowledge, reading and writing abilities, quantitative knowledge and general reasoning capacity.

The verbal analogies test (Hornke & Rettig, 1989) consists of three words. The first two words are set in a relation to one another. Respondents must consider the third word and select a fourth word from a list of options that has the same relation to the third word as the second word did to the first. An example item might be “bird : air = fish : ?”. Respondents must choose the correct word to fill in the question mark out of a selection of

words, in this case “pond”, “spring”, “river” or “water” (Bundeswehr, 2016b). In the present task, “water” would be the right choice. A bird has the same relation to air as a fish has to water: both are the medium in which the animal moves most of the time.

Arithmetic skill is linked to intelligence as well (e.g., Dix & van der Meer, 2015) and is closely connected to various aspects of the CHC model of cognitive abilities, including fluid reasoning, comprehension knowledge, and processing speed (Cormier, Bulut, McGrew, & Singh, 2017).

Hence, arithmetic skill is also tested in the Bundeswehr’s psychological assessment. The arithmetic test consists of different types of mathematical operations. For example: “Three persons need 690 minutes to pave a driveway. How many hours do five persons need? (Result rounded to the next full hour)”[2] (Bundeswehr, 2019a) or “A can of peas costs 0.95€. How much is a box of 42 cans in Euro?”[3] (Bundeswehr, 2019a).

The Bundeswehr’s matrices test (Bundeswehr, 2019a; Hornke, Küppers, & Etzel, 2000) is basically equivalent to the Raven Progressive Matrices (e.g., Raven et al., 2003) in terms of function. Eight simple patterns are represented in a 3 x 3 matrix with the bottom right square left blank. Respondents need to choose the correct missing pattern out of a selection of different options. Visuospatial abilities (visual abilities [Gv] in the CHC model) and reasoning capacity (Gf) strongly influence performance on matrices tests (Waschl, Nettelbeck, & Burns, 2017), which in turn can be seen as strongly loading on these two factors.

[2] The German text is „Drei Personen benötigen für das Pflastern einer Garagenzufahrt 690 Minuten. Wie viele Stunden brauchen fünf Personen dafür? (Das Ergebnis auf ganze Stunden aufgerunden)“

[3] The German text is „Eine Dosen Erbsen kostet 0,95€. Wie viel kostet ein Karton mit 42 Dosen in Euro?“

Online Assessment

Online testing was first conducted in the education sector (Lin, 2011). In this context, the terms e-assessment and OA were used interchangeably (Hertel & Konradt, 2004). OA made it possible to promote online learning and assess new abilities (e.g., Reeves, 2000). As the Internet became more popular and easily accessible for everyone (see Statistisches Bundesamt, 2019 for an overview), companies became interested in using OA for personnel selection purposes (Kupka, Diercks, & Kopping, 2004). In the present case, OA is defined as the assessment of selected abilities in the service of personnel selection.

More than fourteen years ago, almost 10% of companies were already using online pre-employment testing (Piotrowski & Armstrong, 2006). Other organizations of similar size to the Bundeswehr have already been OA for over a decade; for example, Unilever has employed OA since 2004 (Kupka et al., 2004).

OA has several advantages regarding personnel selection and holds a certain appeal: it allows a pre-selection to be made, reducing the costs of the main personnel selection procedure (Galanaki, 2002). Is not restricted with respect to time or place (on-demand testing) and hence quite flexible (Schaper, 2009) and more attractive for applicants (Hertel, Konradt, & Orlikowski, 2003; Kupka, 2013). OA is flexible and saves time in the long run (Barbosa & Garcia, 2005) and can support different kinds of media and self-selection (Schaper, 2009). In addition, OA has the potential to boost tests' quality criteria due to the standardized presentation mode (Jurecka & Hartig, 2007). Computers can reduce measurement and interpretation errors (Ridgeway, McCusker, & Pead, 2004).

In its ideal-typical form, a purely OA approach might look like this: first, applicants conduct a self-assessment to provide a realistic preview of the job. Next,

applicants complete an OA through an online applicant management system. After the OA, an online video interview is conducted (e.g., Schaper, 2009). This approach would rely exclusively on OA. Schaper (2009) illustrates a different process for a prototypical OA, with reference to Bartram (2000) and Hertel et al. (2003). A task analysis is used to determine the core aspects of the job, and the results are used to draft an employment advertisement published online with the option of completing an online application. This represents the online recruiting phase. In the next step, applicants complete tests in an OA for self-assessment purposes, with no information transferred to the employer. If the applicant is still interested, she or he receives a password via email for access to the OA, in which they must complete different tests. Schaper (2009) also mentions that online interviews can be conducted afterwards. Candidates who successfuly complete each of the previous steps are invited to an on-site assessment. To ensure test security, Aguado et al. (2018) suggest a multi-stage procedure in which suspected cheaters are presented with additional test items.

In addition, OA should ideally be very flexible in its application and therefore be able to be used on different electronic devices, such as tablets, laptops, or smartphones. However, it must be taken into account that the medium used could have an influence on performance (for an overview, see Arthur, Keiser, & Doverspike, 2018).

Of course, prognostic validity for job performance is the key to an excellent OA. In general, “selecting out” is recommended for this purpose (Schaper, 2009), meaning that ill-suited applicants do not pass the OA and are not invited to participate in further assessment in order to keep the number of applicants small (Schaper, 2009).

Risks. Although the advantages of OA are clear, it does not come without risk. First, IT literacy and computer skills may have an influence on or even prevent people from completing the OA (Schaper, 2009). However, Albeit, Greiff, Kretzschmar, Müller, Spinath, and Martin (2014) found little evidence for confounding between complex problem solving as measured via computer-based assessment and information and communication technology literacy. Furthermore, since Internet use is now quite common (Statistisches Bundesamt, 2019), this issue may be neglected.

From a purely assessment-oriented perspective, the unstandardized test environment in OA is a big issue (Kantrowitz, Dawson, & Fetzer, 2011). Internet-based ability testing lacks all the monitoring mechanisms typical in computer-based testing, such as motivation of participants (Schroeders, Wilhelm, & Schipolowski, 2010). This is mainly due to the lack of monitoring within OA, also known as UIT (unproctored Internet testing). UIT can be defined as "Internet-based testing of a candidate without a traditional human proctor" (Makransky & Glas, 2011, p. 608). UIT goes along with certain challenges that need to be addressed.

Probably the most pressing issue is the opportunity to cheat, including the risk of testing materials being leaked (Kantrowitz et al., 2011), or applicants taking the test multiple times, which often makes the result of OA unreliable or even invalid. Steger et al. (2018) conducted a meta-analysis of test scores in different testing environments. They found that unproctored assessments are especially vulnerable to cheating: they report a pooled effect of mean differences of $\Delta = 0.20$ (95% CI [0.10, 0.31]). This risk can be decreased by using tests that are hard to search the Internet for ($\Delta = 0.38$, $SE = 0.08$, $p < .001$ vs. $\Delta = 0.02$, $SE = 0.05$, $p = .66$).

Reliability in general seems to be another issue, because the testing mode impacts reliability, validity and acceptance (for an overview, see Konradt, Lehmann, Böhm-Rupprecht, & Hertel, 2003). In ability tests conducted online, it seems that reliability is nearly the same, but performance is worse (Konradt et al., 2003) and test scores are vulnerable to cheating, although the effect is rather small (Kantrowitz & Dainis, 2014; Steger et al., 2018). This effect is larger for speeded tests (e.g., Kurz & Evans, 2004; Potosky & Bobko, 2004; Wilhelm & McKnight, 2002) than for power tests (Mead & Drasgow, 1993). Unfortunately, speeded tests are ostensibly more immune against cheating (Arthur, Glaze, Villado, & Taylor, 2009). However, this finding is not as robust as it seems, since the technical opportunities at the time of the study were limited. Hence, this finding needs to be replicated with current technology.

Flexibility (returning to work on an item after all other items have been finished) seems to result in differences between computers and paper-pencil tests (Bodmann & Robinson, 2004), which was an issue in the early 2000s. A newer study contradicts this finding, showing that the difference in medium (e.g., paper-pencil vs. computer) hardly accounts for individual differences (Schroeders & Wilhelm, 2010). Nevertheless, these results indicate that the test medium should also be closely examined. Therefore, it must be taken into account that different mobile devices can be used in OA and that this has an impact on performance, especially in cognitive testing (Arthur, Doverspike, Muñoz, Taylor, & Carr, 2014). However, while the display size seems to have no influence if only computers are considered (Chen & Perie, 2018), the nature of the test may have an impact (Bridgeman, Lennon, & Jackenthal, 2003).

Another complication of OA is that no one can explain the instructions, and it is difficult to ascertain whether they were understood correctly (Wilhelm & McKnight,

2002). Nevertheless, all of these aforementioned quality challenges can be reduced to a minimum if the test is developed specifically for OA (Schaper, 2009).

The issue of cheating remains problematic, however. According to Arthur et al. (2009), there is not much research on cheating in employment testing, even though the wide prevalence of cheating in educational settings clearly indicates that cheating is a major issue in UIT. There has been some research on UIT performance in low-stakes testing (Domínguez et al., 2019), but evidence for high-stakes testing is mostly lacking. The few available results indicate that while there is some cheating, it does not seem to occur very often (Aguado et al., 2018; Kantrowitz & Dainis, 2014). However, there is a considerable amount of research concerning faking in questionnaires or faking in interviews (e.g., Bensch, Maaß, Greiff, Horstmann, & Ziegler, 2019; Pelt, van der Linden, & Born, 2018; Roulin & Powell, 2018), which is considered as both a continuous and quantitative variable (Ziegler, Maaß, Griffith, & Gammon, 2015). Unfortunately, faking seems to be an issue in the military context as well (Boss, König, & Melchers, 2015). Moreover, these results indicate the relevance of cheating in online cognitive ability assessments (Carstairs & Myors, 2009; Cavanagh, 2014). Hence, major issues for the development of an OA are to reduce cheating (e.g., Steger et al., 2018) and make the assessment reliable, since it seems to be the most prominent risk. There are also a few technical problems to be considered (e.g., data security, server capacity or transmission rate), but these will not be addressed here (see Schaper, 2009 for further information).

Possible solutions. Before listing the different ways to circumvent the risks of OA, the bad news first: No matter which approach is taken, the biggest risk, namely cheating, may be detected, but can never be eliminated entirely (Schroeders et al., 2010). However, this could be said about any computer-based testing environment with many applicants well. Various approaches can be chosen to detect cheating, like response time, unusual response patterns (van der Linden & van Krimpen-Stoop, 2003), adaptive tests using a likelihood ratio or adaptive confirmation testing (Makransky & Glas, 2011). Guo and Drasgow (2010) recommend an additional test to detect inconsistent test results, which seems to work well in practice (Aguado et al., 2018).

Wiedmann (2009) suggests a different approach, namely telling applicants that they will be retested in a controlled setting, making clear that any attempt to cheat would come to light in this subsequent assessment. The International Test Commission (2005) takes yet another approach, recommending the following:

> For moderate and high stakes assessment (e.g., job recruitment and selection), where individuals are permitted to take a test in controlled mode (i.e. at their convenience in nonsecure locations), those obtaining qualifying scores should be required to take a supervised test to confirm their scores.
>
> - Procedures should be used to check whether the test-taker's original responses are consistent with the responses from the confirmation test.
> - Test-takers should be informed in advance of these procedures and asked to confirm that they will complete the tests according to instructions given (e.g., not seek assistance, not collude with others etc.).

> This agreement may be represented in the form of an explicit honesty policy which the test-taker is required to accept. (International Test Commission, 2005, p. 33)

However, while there is evidence that such honesty policies produce negative reactions in personality measures (Converse et al., 2008), similar evidence seems to be lacking for cognitive ability testing.

Other approaches include a registration code which can only be used once in order to reduce the risk of repeating the test or letting another person complete it, because each participant has only chance to succeed (Bartram, 2000). Another recommendation is to vary the order of the items and not allow participants to go back and forth among items, i.e. through parallel versions of the test or a randomized item order. This can make it more difficult for groups to work on the test together (Schaper, 2009). For obvious reasons, only tests that are quite hard to cheat on should be used, unlike, for example, arithmetic tests, which can easily be cheated on with a calculator.

These approaches need not be applied in isolation, but can be combined. For example, a massive item pool combined with a randomized item design would be an option to reduce the risk of cheating due to leaked test material. The advantage of this approach is that it will take a long time before all items are known, and even if this occurs, it will probably be hard for applicants to remember which answer belongs to which item. Chen, Lei, and Liao (2008) recommend a combination of item exposure control and test overlap to ensure high test safety in adaptive testing, which in turn creates a need for many different items. This is associated with substantial cost and time for item production and calibration. Rudner (2010) assumes that it takes 1,500 to 2,500 US dollars to create one item using traditional approaches and testing procedures. Therefore, a massive item pool is

quite costly. The costs are even higher for OA, in which many items are required to ensure test safety, regardless of test mode.

Another approach would be item generation on the fly, meaning that an individual item set is created for each participant the moment he or she starts the test. This approach can be applied either by randomly drawing items from an item pool, as described above, or through automatic item generation. Due to its complexity, the latter is described in the section on automatic item generation.

Working Memory

WM is essential in all domains of everyday life and cognitive activities (Engle, 2002). Although definitions of WM vary (Cowan, 2017), the research seems to agree on what function WM represents. It can be defined as "the mechanisms and processes that hold the mental representations currently most needed for an ongoing cognitive task available for processing" (Oberauer, 2019b, p. 1), or more succinctly, "working memory capacity is simply the ability to remember things in an immediate-memory task." (Cowan, 2005, p. 2). Although working memory capacity (WMC) has been discussed for quite some time (Miller, 1956), and measured even longer (Terman, 1916), with the term itself introduced in 1960 (Miller, Galanter, & Pribram, 1960), it was not until 1968 that Atkinson and Shiffrin described a "short-term store" that could be seen as WM and not until 1974 (Baddeley & Hitch, 1974) that a holistic model was introduced. The essential role of WM became obvious early, as it plays an important role in performance (Daneman & Carpenter, 1980) and in predicting a wide range of cognitive abilities (Kane, Conway, Hambrick, & Engle, 2007). It is defined as a multicomponent system (Kane, Conway, Hambrick et al., 2007), which is why capacity differs between individuals and in turn why

the executive functions are more efficient. Consequently, is it is not surprising that WM is closely related to other cognitive functions. Engle, Kane, and Tuholski (1999) argue that WMC is more about controlled attention than remembering and storing information. They posited a close relation between attention and WM, which has been confirmed empirically (Baddeley & Logie, 1999). Baddeley (1993) claimed that WM refers to the capacity to distribute attention rather the control mechanism.

Despite the absense of a discussion in the literature about what functions define WM, different models of the concept have proposed. The most popular models are described below.

Working memory models. As already mentioned, the first holistic WM model was proposed by Baddeley and Hitch (1974) and further developed over the years (Baddeley, 1986a, 2000). It can be considered the basis for most research regarding WM (Dehn, 2015). Baddeley and Hitch (1974) demonstrated that stimuli coded in the same way are more difficult to process than differently coded stimuli, leading them to the conclusion that WM entails different storage systems (short-term memory) and a processing component. Their model consisted of four parts. The two storage systems are the phonological loop and the visuospatial sketchpad, which are considered "slave systems" to the processing component. The phonological loop holds speech-based information that is actively repeated (rehearsal process) (e.g., Awh et al., 1996; Baddeley, 2003). The visuospatial sketchpad holds visual and spatial information (Baddeley, Grant, Wight, & Thomson, 1975; Baddeley & Lieberman, 1980; Logie, 1995; Repovš & Baddeley, 2006). A third component, the episodic buffer, was added later on (Baddeley, 2000). This component is able to encode in a multimodal way and is limited as well. The fourth and

probably most important component is the executive control system ("central executive", Baddeley, 2002). This system controls and supervises the three other components and attention (e.g., Cowan, 1999; Engle, Tuholski, Laughlin, & Conway, 1999). Therefore, it is independent but able to interact with the other components. In addition, the central executive is assumed to be responsible for knowledge transfer from long-term memory (LTM) to WM (Repovš & Baddeley, 2006). In summary, then, the central executive is considered to be responsible for four processes: focusing attention, divided attention, switching focus, and the retrieval and integration of LTM and WM (Baddeley, 1996, 2007). However, the attention functions of the central executive seem to be most crucial (Baddeley & Logie, 1999), explaining differences in WM (e.g., Engle & Kane, 2004; Engle, Tuholski et al., 1999; Kane, Bleckley, Conway, & Engle, 2001; Kane, Conway, Hambrick et al., 2007).

In another line of research, Oberauer (2009) extended the WM model by Cowan (1997; Cowan, 1999) based on research by Cowan (1988) and himself (Oberauer, 2002). Oberauer's model emphasizes attention (e.g., Oberauer, 2019b), which is a limited resource and therefore limits WM, although this explanation is not without flaws (Oberauer, Farrell, Jarrold, & Lewandowsky, 2016). In this model, WM itself can be seen as a form of attention (Oberauer, 2019b).

Oberauer (2009) states six requirements of WM: structural representation, manipulation, flexible reconfiguration, partial decoupling from LTM, retrieval from LTM and encoding structural information into LTM. This model distinguishes between declarative and procedural WM. Declarative WM is responsible for representing content, and procedural WM for processing (Oberauer, 2009). These two components can be compared to Baddeley's central executive and slave systems (Baddeley, 1986b; Oberauer, 2009). The declarative component consists of three parts: the activated part of LTM, the

region of direct access and the focus of attention. Two statements can be made about the activated part of LTM. First, the time for retrieval decreases as activation rises, and second, content similar to already activated content can be processed more quickly. Stimuli in the direct access region are smaller in number and can be processed in structures and reference systems, similar to chunking. This ability is necessary for inductive reasoning (Oberauer, 2009). As the region of direct access has a limited capacity (Oberauer, 2009), this component can be understood as WMC (Oberauer, 2005b). The focus of attention can access and manipulate the stimuli held in the direct access region. The procedural component of WM has access to the direct access region and therefore includes the focus of attention. In a more recent paper, Oberauer (2019b) specifies the relation between attention and WM and makes five claims that are supported by previous research: First, WM is a form of attention. Second, the information held in WM is a form of controlled attention. Third, paying attention to an object does not guarantee that it will be encoded in WM. Fourth, the focus of attention can be shifted and the selected items manipulated. Fifth, like templates, representations in WM guide and influence the control of attention and action.

It remains unclear whether WM encompasses two distinct factors, namely a spatial and a non-spatial factor (Oberauer, Süß, Wilhelm, & Wittmann, 2003), as has been previously claimed in the literature (Daneman & Tardif, 1987; Kyllonen & Christal, 1990; McCants, Katus, & Eimer, 2018; Oberauer, Süß, Schulze, Wilhelm, & Wittmann, 2000; Shah & Miyake, 1996; Smith & Jonides, 1997).

The CHC model of cognitive abilities (Schneider & McGrew, 2012, 2018), based on Raymond Cattell, John Horn and John Carroll's psychometric approach to intelligence theory represents a more holistic approach to WM. This theory can be depicted as a hierarchy, as seen in Figure 1. The broadest ability is *g,* which can be divided into several

different broader abilities. Each broader ability is associated with narrow abilities, which are in turn associated with specific abilities that can be measured with specific tests. The specific abilities represent the bottom of the hierarchy. The broader abilities can be clustered into categories. Gf is domain-general reasoning capacity and therefore stands by itself. Other abilities can be clustered under acquired knowledge abilities, namely comprehension knowledge (Gco), domain-specific knowledge (Gkn), reading and writing abilities (Gw) and quantitative knowledge (Gq). Another cluster covers domain-specific sensory abilities like visual abilities (Gv), auditory abilities (Ga), olfactory abilities (Go), tactile abilities (Gh), kinesthetic abilities (Gk) and psychomotor abilities (Gp).

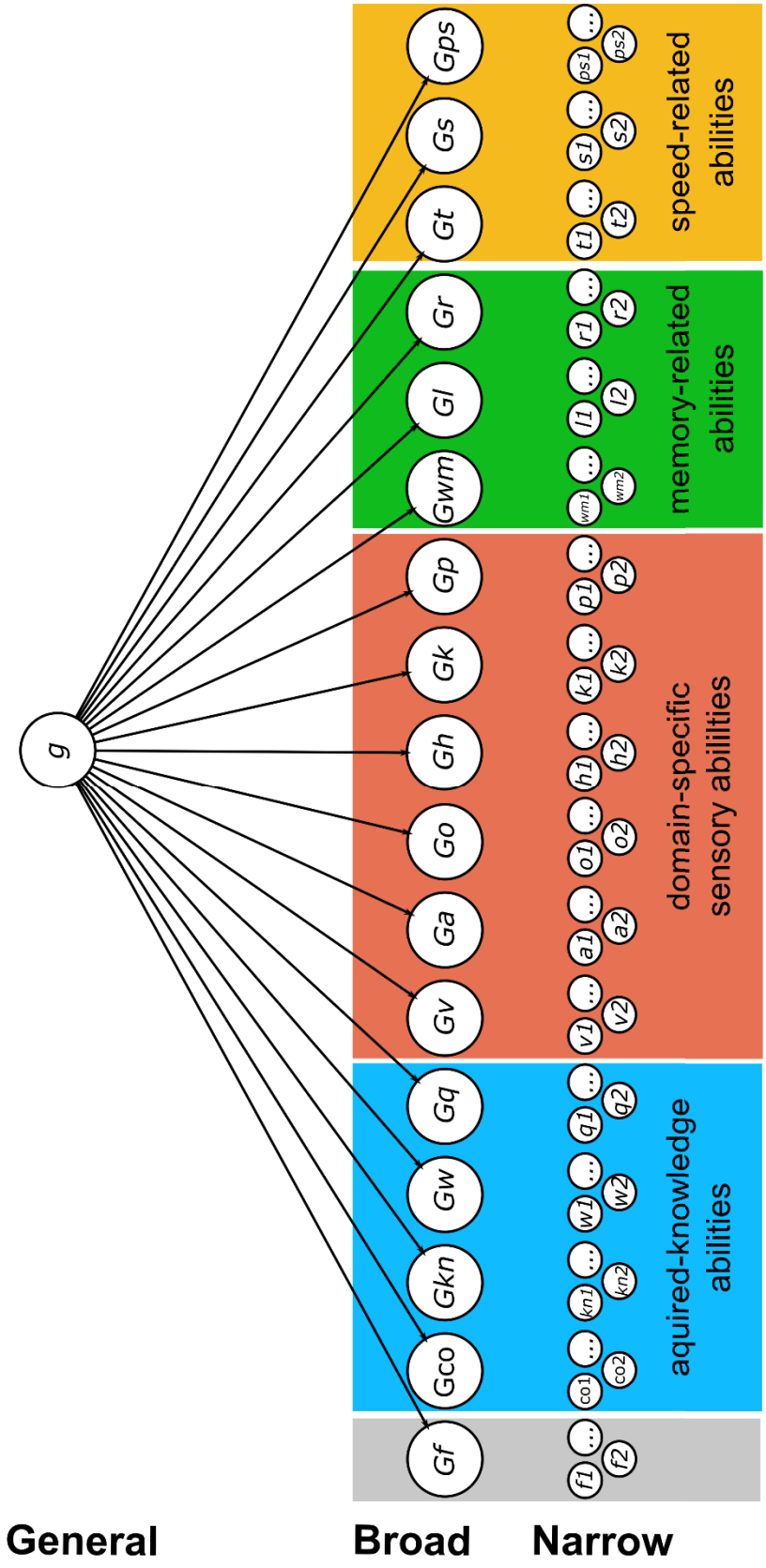


Figure 1. Hierarchical depiction of CHC model.

Yet another cluster encompasses memory-related abilities like WMC (Gwm), learning efficiency (Gl), and retrieval fluency (Gr). The final cluster entails speed-related abilities like reaction/decision time (Gt), processing speed (Gs), and psychomotor speed (Gps). While Schneider and McGrew (2018) describe these clusters in their CHC theory, they also report a conceptual grouping of abilities. In this grouping, *motor abilities* encompasse psychomotor abilities and psychomotor speed, while *perceptual processing* encompasses all domain-specific sensory abilities except psychomotor abilities and reaction/decision time. *Controlled attention* encompasses fluid reasoning, WMC and processing speed. The last group is *acquired knowledge,* which encompasses quantitative knowledge, domain-specific knowledge, comprehension knowledge, retrieval fluency and learning efficiency. As can be seen, CHC theory embeds WM in a broader spectrum of abilities.

Since many WM models now exist, as described above, efforts are currently being made to at least set benchmarks for WM models (Oberauer et al., 2018). Thus, it can be assumed that the development of WM models is far from complete.

Working memory capacity. The question of WMC is not easily answered. Originally, Miller (1956) proposed a 7±2 capacity limit on WM, defining a unit as a "bit of information" (Miller, 1956, p. 83). He explains deviations from this capacity limit through chunking, where a chunk can hold an extended bit of information without violating the proposed limit. However, it does not seem to be that easy. Cowan (2000) proposes a limit of 3 to 5 chunks, if no strategy is applied. This is in accordance with the debate concerning a capacity limit of four different chunks in the literature (Awh, Barton, & Vogel, 2007; Luck & Vogel, 1997; Olson & Jiang, 2002). However, studies differ in their answers to the

question of whether the capacity limit is only four (Cowan, 2000; Miller, 1956) or four for each subfacet of WM. For instance, visual WM is limited to four (Alvarez & Cavanagh, 2004; Awh et al., 2007; Davis & Holmes, 2005; Luck & Vogel, 1997; Olson & Jiang, 2002), whereas studies indicate that WMC for lists of digits can be trained and reach up to 90 (e.g., Kliegl, Smith, Heckhausen, & Baltes, 1987). Although interference seems to work the same way in verbal and visual WM (Oberauer & Lin, 2017), it seems that relations between objects can boost WMC (O'Donnell, Clement, & Brockmole, 2018), which would fit with Oberauer's (2019c) insight that not items but bindings are relevant for WMC. However, visual WM seems to be particularly vulnerable to interference (Makovski, 2016; Shoval, Luria, & Makovski, 2019). Nassar, Helmers, and Frank (2018) attribute this to strategies focusing on accuracy or capacity in WM, which is influenced by chunking.

Therefore, strategies and chunking must be considered separately. Although WM seems to be a domain-general construct, in contrast to short-term memory (Kane et al., 2004), more verbal cues can be remembered better due to chunking (Cowan, Rouder, Blume, & Saults, 2012). Cowan (2000) defines a chunk "as a collection of concepts that have strong associations to one another and much weaker associations to other chunks concurrently in use"[4] (Cowan, 2000, p. 89). The application of a strategy or chunking makes it easier to keep a large amount of information active (Chase & Simon, 1973). Thalmann, Souza, and Oberauer (2019) conclude that chunking relieves WM by retrieving a compact chunk from LTM instead of needing to remeber the individual items.

Normally, when measuring visual WM, mechanisms are implemented that prevent participants from applying a strategy like verbal encoding (Luck & Vogel, 1997; Makovski & Jiang, 2008), suggesting that verbal encoding may boost performance. In addition, proactive interference contributes to differences in both verbal WM (Kane & Engle, 2000)

[4] Cowan (2000) uses the term "concepts" to refer to concepts in LTM.

and visual WM (Hartshorne, 2008; Makovski & Jiang, 2008; Postle, Brush, & Nick, 2004). Proactive interference is interference from a previous trial negatively impacting performance on a subsequent trial (Baddeley, 1990; Keppel & Underwood, 1962). "Item-specific PI [proactive interference] occurs when the response-eliciting probe on the current trial matches an item not from the current memory set […], but from the memory set of the previous trial" (Postle & Brush, 2004; see also Wickens, Born, & Allen, 1963). Item-nonspecific proactive interference is produced by the test length, because more items are displayed (Keppel & Underwood, 1962; Postle & Brush, 2004; Wickens et al., 1963; Wickens, 1973). Unsurprisingly, proactive interference influences performance in WM (Kane & Engle, 2000) and WM span tasks (Lustig, May, & Hasher, 2001). Proactive interference seems to play a crucial role in most WM tasks due to the similarity of the stimulus material and test length, making it difficult to determine WMC. In addition, retroactive interference (Rosen & Engle, 1998) plays an important role as well. Interference control and WMC are closely related, as are interference control and fluid intelligence (Unsworth, 2010). Therefore, interference should not necessarily be excluded when determining WMC (Xu, Adam, Fang, & Vogel, 2018).

One reason for the role of interference in WMC could be that content in WM must be removed selectively and item-specifically (Ecker, Oberauer, & Lewandowsky, 2014). Lewis-Peacock, Kessler, and Oberauer (2018) postulate that content can be deleted from WM in two ways: one is only temporary and does not alter the link between content and context, while the other is irreversible and releases the bind between content and context. In the first case, interference could cause a problem because removal it is reversible and old content blocks new. Moreover, only the irreversible removal of information releases WMC (Oberauer, 2018), which is needed to store new information. While Farrell et al. (2016) state "we should note the possibility that forgetting from WM is not solely due to

decay (counteracted by rehearsal) or interference (counteracted by distractor removal), but that both processes might contribute to forgetting" (Farrell et al., 2016), other researchers suggest completely rejecting the decay hypothesis (Oberauer et al., 2016), making interference an even more integral part in WMC.

Cowan (2010) concludes that the determination of WMC is quite difficult due to the multiple mechanisms involved.

Working memory tests. There are a broad variety of tests measuring WM (e.g., Case, Kurland, & Goldberg, 1982; Daneman & Carpenter, 1980; Oberauer et al., 2000; Salthouse & Mitchell, 1989; Süß, Oberauer, Wittmann, Wilhelm, & Schulze, 2002; Turner & Engle, 1989; Wechsler, 2008). Despite some tests being seemingly unfit to measure WM, such as the n-back task (Jaeggi, Buschkuehl, Perrig, & Meier, 2010; Kane, Conway, Miura, & Colflesh, 2007; Szmalec, Verbruggen, Vandierendonck, & Kemps, 2011), they are widely used in fields such as neurology (e.g., Cui, Bray, Bryant, Glover, & Reiss, 2011; Haberecht et al., 2001; Hoeft et al., 2007; Kesler et al., 2004). This makes it necessary to take a closer look at measurements of WMC. Oberauer et al. (2000) give a good overview of different kinds of tests. They distinguish among three categories of WM tests: coordination tasks, supervision tasks and storage and transformation tasks. For example, a reading span would fall under all three categories, whereas a backward digit span (DS) task can be categorized as storage and transformation, and updating tasks can be categorized as storage and transformation and coordination (Oberauer et al., 2000). Span tasks in general are quite popular. They necessitate not only storage and processing but also simultaneously processing additional information (Case et al., 1982; Daneman & Carpenter, 1980; Turner & Engle, 1989). They are also reliable and valid (Conway et

al., 2005). Performance in span tasks is influenced by "multiple factors, with domain-specific skills, such as chunking and rehearsal, facilitating storage and a domain general capability allowing for cognitive control and executive attention" (Conway et al., 2005, p. 771). However, it seems that rehearsal has little effect on WM performance (Oberauer, 2019a). Furthermore, span tasks are quite easy to explain and conduct because the test design is quite simple.

Both complex and simple span tasks exist (e.g., Redick & Lindsey, 2013). Simple span tasks are more likely to measure short-term memory due to the lack of a processing component (Conway, Cowan, Bunting, Therriault, & Minkoff, Scott, R. B., 2002; Conway & Engle, 1996; Daneman & Carpenter, 1980; Dixon, LeFevre, & Twilley, 1988; Kail & Hall, 2001; Turner & Engle, 1989) and therefore seem to measure only part of WM (e.g., Baddeley & Hitch, 1974). However, researchers reanalyzing prior studies found evidence against this hypothesis (Colom, Rebollo, Abad, & Shih, 2006; Unsworth & Engle, 2007). Therefore, no clear conclusion on which span tasks are better can be drawn, although Schmiedek, Hildebrandt, Lövdén, Wilhelm, and Lindenberger (2009) clearly state that only complex span tasks can measure WMC equally well as updating tasks. Because complex span tasks meet the requirement that WM measures should capture the simultaneous processing of additional information (Case et al., 1982; Daneman & Carpenter, 1980; Turner & Engle, 1989), they can be perceived as a measure of executive WM (Dehn, 2015). Cowan et al. (2005) propose that "the critical aspect of successful WM measures is that rehearsal and grouping processes are prevented, allowing a clearer estimate of how many separate chunks of information the focus of attention circumscribes at once" (p. 42), which seems to be possible in both types of span tasks depending on the stimulus material. In addition, it is worth mentioning that complex span tasks account for half the variance in measures of general fluid abilities (Kane et al., 2005),

which is more than simple span tasks explain (e.g., Conway et al., 2002; Engle, Tuholski et al., 1999). However, short-term memory accounts for the relation between reasoning and WM (Krumm et al., 2009) and seems to explain the relation between WM and fluid intelligence in children (Hornung, Brunner, Reuter, & Martin, 2011).

Scoring. Which scoring procedure is applied depends heavily on the test, which is why this chapter focuses only on span tasks. Conway et al. (2005) provides a good overview of scoring methods for span tasks. They differentiate among four scoring methods: partial-credit unit scoring, all-or-nothing unit scoring, partial-credit load scoring and all-or-nothing load scoring. In partial-credit unit scoring, a total score of 1 can be reached for each item. 1 is given if the answer is completely correct. All scores between 0 and 1 are possible for an item. For example, if 2 out of 4 stimuli are recalled correctly, the score would be 0.5. At the end of the test, the scores for each item are added up. All-or-nothing unit scoring can be considered the classical approach, where each correctly answered item is given a score of 1. Even if just one stimulus was answered incorrectly, a score of 0 is given for the item. Partial-credit load scoring is similar to all-or-nothing unit scoring, except that the number of stimuli is taken into account. For example, for an item with four stimuli and three correct responses, a score of 3 would be awarded. For an item with three stimuli and all three answered correctly, the score is 3 as well. All-or-nothing load scoring likewise takes the number of stimuli into consideration. However, similar to all-or-nothing unit scoring, only correctly answered items are considered. For better understanding, an example can be seen in Table 1. Fictitious items from a DS backwards task are depicted in the first column, followed by fictitious responses in the second column. The subsequent columns show the test scores as calculated by the different scoring procedures. At the bottom of the table, the fictitious participant's score in relation

to a would-be full score followed by the score in percent are listed to provide a deeper understanding.

Table 1

Examples of different scoring methods

DSB item	Fictional answer	Scoring Procedure			
		All-or-nothing unit scoring	Partial-credit unit scoring	All-or-nothing load scoring	Partial-credit load scoring
5 9 7	7 9 5	1	1	3	3
1 7 6 2	2 6 7 2	0	0.75	0	3
3 5 4 2 6	5 5 4 5 3	0	0.60	0	3
2 7 9 4 1 6	6 1 4 9 7 2	1	1	6	6
9 2 4 3 7 1 8	8 1 7 3 4 2 9	1	1	7	7
score in relation to full score		3/5	4.35/5	16/25	22/25
score in percent		60%	87%	64%	88%

Note. DSB – DS backward.

The evidence for weighted vs. non-weighted scoring is not so strong. However, Conway et al. (2005) recommend weighted scoring and thus partial-credit load scoring, as it leads to greater differentation than non-weighted scoring (Bensch et al., 2019). This type of scoring seems to make sense at least for visual WM, which is object-based for difficult tasks (Qian, Zhang, Liu, & Lei, 2019).

Working memory in personnel selection. The use of WM in personnel selection is not new. Fifteen years ago, it was proposed that WM could be relevant for personnel selection (Colom, Martínez-Molina, Shih, & Santacreu, 2010; König, Bühner, & Mürling, 2005). Edwards, Franco Watkins, McAbee, and Faura (2017) cited three reasons why WM should be highly relevant for job performance: first, WM is important for learning new skills; second, WM is important for reproducing previously learned content; and third, WM helps suppress irrelevant information while keeping relevant information active.

Furthermore, intelligence is the best predictor of job performance if only one predictor is used (Kuncel, Hezlett, & Ones, 2004; Ree et al., 1994; Ree & Earles, 1992; Schmidt, Hunter, & Outerbridge, 1986). Therefore, intelligence assessment is often part of professional personnel selection (for an overview see Schmidt & Hunter, 2000). As the association between WM and fluid intelligence accounts for the main differences in *g* (Ackerman, Beier, & Boyle, 2002; Cowan, 2005; Kyllonen, 1996), it is unsurprising that WM and *g* are closely connected (Gignac, 2014; Gignac & Watkins, 2015; Kane et al., 2005; Oberauer et al., 2005). Giofrè, Mammarella, and Cornoldi (2013) could even replicate the finding that WM accounts for a huge proportion of *g,* up to 66%. Other studies come to the same conclusion: Wittmann and Süß (1999) stated that general intelligence and WM shared 53.5% of common variance (measured with the BIS, Jäger, Süß, & Beauducel, 1997). This is not surprising in light of the CHC theory of intelligence (Schneider & McGrew, 2012, 2018), in which WM is a subordinate ability to *g*. It has also been found that WM and fluid intelligence are very closely related (Rey-Mermet, Gade, Souza, Bastian, & Oberauer, 2019). The parts of WM responsible for this connection are the executive functions and not, for example, short-term memory (Dehn, 2015). Alloway and Alloway (2010) even suggest that WM is a more important predictor of academic

performance than intelligence measured with the Wechsler Objective Reading Dimensions (Wechsler, 1993) and Wechsler Objective Numerical Dimensions (Wechsler, 1996).

Although there is limited evidence of civilian companies using WM for personnel selection and even less evidence for the predictive validity of WM for job performance, it is only logical to consider the measurement of WM for that purpose. Moreover, various countries' militiary already use WM for personnel selection purposes, including the Canadian military (Kemp & Jalbert, 2012), Swedish military (Wolgers, 2015) and British military (Irvine, 2014).

Test Theory

Test theory should serve as the basis of every psychological test, describing the relation between the ability being measured and test responses (Rost, 2004). Two types of test theories exist: classical test theory (CTT) and item response theory (IRT). CTT is the older theory of the two, going back to times where complex calculations could not be conducted easily (Gulliksen, 1950; Lord & Novick, 1968; Zimmerman, 1975). However, CTT has several issues that IRT tries to solve. The next section describes IRT only. For an overview of both theories, see Rost (2004), Bühner (2011) or Kline (2016).

Item response theory. IRT can be traced back to a model by Rasch (1960) (the Rasch model [RM] or 1PL model). IRT is not a single theory, but rather a family of probabilistic models (Rost & Spada, 1982). In IRT, the person parameter (θ) and item parameter (β) are on the same scale (Hambleton & Slater, 1997), can be estimated separately (Rasch, 1960) and are largely not dependent on the sample (Hambleton, Swaminathan, & Rogers, 1991). The item parameter is estimated by looking for the best

possible match between person ability parameters and item difficulty parameters. One method for accomplishing this goal is likelihood estimation, in which the maximum likelihood for estimating the item parameter and person parameter (Kubinger, 2019) is sought for a given dataset. Joint maximum likelihood (Wright & Panchapakesan, 1969) is a method of determining maximum likelihood developd quite early. However, estimations with joint maximum likelihood are not stable, and the problem increases with sample size (see e.g., Haberman, 1977). Thus, a more popular approach is marginal maximum likelihood estimation (Bock & Aitkin, 1981), where the estimations are stable. The final approach mentioned here is conditional maximum likelihood (Mair & Hatzinger, 2007), which is as least as good as marginal maximum likelihood (Mair & Hatzinger, 2007). However, zero and perfect scores are taken into consideration in marginal maximum likelihood estimation but not in conditional maximum likelihood, since the formula cancels the person parameter out (Mair & Hatzinger, 2007). If the person parameter is not of concern but rather the item parameter difficulty, conditional maximum likelihood seems to be the better approach, since zero and perfect scores tend to be unreliable because the estimation can approach ± infinity (for a comparison, see Mair & Hatzinger, 2007)[5].

The following sections describe specific IRT models within the Rasch family.

[5] In addition, the following likelihood methods should be mentioned for the sake of completeness: the pseudo-maximum likelihood approach by Anderson, Li, and Vermunt (2007), non-maximum likelihood approaches by Molenaar (1995) and Linacre (2004) and a Bayesian approach by Baker and Kim (2004).

General properties of Rasch models. The family of RMs share the following assumptions: unidimensionality (e.g., Glas & Verhelst, 1995) of the latent trait or homogeneity (Hattie, 1985; Pedhazur & Schmelkin, 1991), which means that answers to the test items are mainly determined by the ability being measured and nothing else. Expressed differently: "unidimensionality is defined as the existence of one latent trait underlying the data" (Hattie, 1985, p. 139). Unidimensionality can be tested with Andersen's likelihood ratio test ([LRT], Andersen, 1973; Glas & Verhelst, 1995). An indicator for item homogeneity is inter-item correlation, which can be evaluated by a factor analysis, for example (Lord & Novick, 1968). Furthermore, items must be locally independent. This means that the probability of correctly solving an item i for a person v with θ_v is only dependent on the ability parameter and not the probability of solving another item. A too-high correlation between items after the contribution of the latent trait (the trait the items are supposed to measure) is removed indicates that this assumption has been violated (Lee, 2004). If items are locally independent, inter-item correlations can be explained by the ability parameter of a person v (θ_v). In this context, "local" refers to the person's ability value, meaning the solution probability is examined with respect to a constant ability parameter. In addition, sample independence is important to support the claim that the results are independent of the sample. This includes subgroup invariance, meaning that parameter estimation does not differ across subgroups (e.g., men and women). Another assumption is sufficiency of the raw scores, which means that raw scores sufficiently reflect persons' ability (Rasch, 1960). Although it was long discussed whether this assumption holds for polytomous RMs (see Andrich, 2016 for an overview), a simulation study could show that raw score sufficiency is indeed applicable to polytomous RMs as well (Andrich, 2010). However, Rost (2001) questions whether these properties hold for all RMs, such as linear logistic test models (LLTM). The last but not least

prerequisite are parallel item characteristic curves (e.g., Mair & Hatzinger, 2007), meaning that the solution probability increases as a person's ability increases (e.g., Koller, Alexandrowicz, & Hatzinger, 2012).

Nevertheless, there is no clear indication whether violations of these assumptions are a problem in RMs. Some studies suggest this is not problematic (Anderson, Kahn, & Tindal, 2017; Dorans & Kingston, 1985; Guilleux, Blanchin, Hardouin, & Sébille, 2014; Rentz & Bashaw, 1977), while others say it is (Brandt, 2012; Loyd & Hoover, 1980; Skaggs & Lissitz, 1986; Slinde & Linn, 1978). When the prerequisite of local independence is violated, it is recommended to use non-parametric IRT instead of parametric, although the parameters obtained from both methods are quite similar (Dirlik, 2019; Meijer, Sijtsma, & Smid, 1990). However, most items have both major and minor dimensions and are therefore not truly unidimensional (Bolt & Lall, 2003; Nandakumar, 1991). "An item is considered unidimensional if the systematic differences within the item variance are only due to one variance source, that is, one latent variable." (Ziegler & Hagemann, 2015, p. 231). Therefore, it remains unclear whether and to what extent violations of assumptions bias parameter estimation.

All of the models below can be described as RMs or extended RMs (Mair & Hatzinger, 2007). Since all of the models share common properties and are built similarly, they can be put into a hierarchy (Mair & Hatzinger, 2007): the most general model is the linear partial credit model (LPCM), followed by its derivate the partial credit model (PCM). The LLTM is further on in the hierarchy, while the most specific model is

the RM[6]. The hierarchical order of the models follows their chronological development and logical deduction.

Rasch model. The RM (Rasch, 1960) can be described with

$$P\ (X_{vi} = x_{vi}\ |\theta_v, \beta_i\ ) = \frac{\exp(0(\theta_v - \beta_i)) * \exp(1(\theta_v - \beta_i))}{1 + \exp(\theta_v - \beta_i)}$$

where P is the probability of an answer x_{vi}, given the ability parameter θ and the item difficulty β. Expressed differently, P is the probability of correctly solving an item i with $i = 1, \dots, k$ with item difficulty parameter β for a person v with $v = 1, \dots, n$ with an ability parameter θ. This can also be expressed as p_{vi}. For example, if the item difficulty parameter and ability parameter are the same, a person v has a 50% chance ($p_{vi} = 0.5$) of correctly answer the item i[7]. Items need to be binary to fit the RM (Koller et al., 2012; Mair & Hatzinger, 2007). The first expression depicts the probability of not correctly answering the item (answer = 0), while the second expression depicts the probability of correctly answering the item (answer = 1). The formula can be shortened and reformulated as

$$P\ (X_{vi} = 1\ |\theta_v, \beta_i\ ) = \frac{\exp(\theta_v - \beta_i)}{1 + \exp(\theta_v - \beta_i)}$$

This means that the correct answer to an item is dependent on the person's ability parameter and the item difficulty parameter. In the conditional maximum likelihood approach, participants who correctly solved all or no tasks need to be excluded (Mair

[6] For the sake of completeness, the linear rating scale model (LRSM) and the rating scale model (RSM) should also be mentioned here. In the hierarchy, the LRMS comes after the LPCM and before the LLTM. The RSM is a derivation of the LRSM, as Mair and Hatzinger (2007) describe.

[7] Because exp (0) = 1

& Hatzinger, 2007) because their ability parameter cannot be estimated; it approaches plus or minus infinity.

Partial credit model. The PCM was proposed by Masters (1982) and can be expressed with

$$P(X_{vih} = 1) = \frac{\exp(h\theta_v - \beta_{ih})}{\sum_{l=0}^{m_i} \exp(l\theta_v - \beta_{il})}$$

It is quite similar to the RM with the exception that items do not need to be dichotomous and scoring need not be constant across items. In the formula, the categories, represented with h, reflect this property. Therefore, each β_{ih} describes an item-category combination in the PCM. If all categories are answered correctly, the item is answered correctly as a whole, but it is possible to get partial points for partial correctly answered item parts (e.g., answering three out of five terms correctly in the DS).

Linear logistic test model. The LLTM proposed by Fischer (1973) is a special case of a generalized linear model (DeBoeck & Wilson, 2004). It is equivalent to the RM, but the item difficulty parameter is linearized as follows:

$$\beta_i = \sum_{j=1}^{p} w_{ij}\eta_j$$

where β_i is the item difficulty parameter for item i which consists of the sum of all difficulty parameters η of category j with weights w_{ij} for the respective item and category, which are determined in advance via a design matrix w (k x p with item i with $i = 1, \dots, k$ and categories j with $j = 1, \dots, p$).

Hence, it can be shortened to

$$\beta = W\eta$$

Linear partial credit model. The LPCM (Fischer & Ponocny, 1994) is the linearized model form of the PCM. The item difficulty parameter for each category can be reformulated into the linear combination

$$\beta_{ic} = \sum_{j=1}^{p} w_{icj}\eta_j$$

Where β_i is the item difficulty parameter for item i with item category c. It consists of the sum of all difficulty parameters η of category j with weights w_{ij} and item category c for the respective item. As in the LLTM, the design matrix W (o x p with $c = 1, \ldots, o$ and $j = 1, \ldots, p$) needs to be defined a priori.

Automatic Item Generation

The basic notion of automatic item generation was first suggested quite early (Bormuth, 1969). Although Fischer (1973) applied the method for the first time a few years later, only later did it become a focus of research and was applied to different areas of psychometrics. Irvine and Kyllonen (2002) or Gierl and Haladyna (2012) give good overviews of the many possible uses of automatic item generation.

Automatic item generation for military use was introduced quite early (Irvine, Dann, & Anderson, 1990; Kyllonen, 2003) and has even previously been applied in the German military (Goeters & Lorenz, 2002). Nowadays, a greater variety of tests are based on automatic item generation (e.g., Arendasy & Sommer, 2010; Arendasy, Sommer, &

Mayr, 2011; Baghaei & Ravand, 2015; Bejar, 1990; Embretson, 1984; Gierl & Lai, 2018), and it is used in different test modes, such as situational judgment (Bejar & Cooper, 2013) or multiple-choice tests of medical knowledge (Gierl, Lai, & Turner, 2012).

There are several advantages of automatic item generation. First, because there is a basic template, items can easily be manipulated and many items can be created (Gierl & Lai, 2016). Second, because there is no need to calibrate each item, it is a lot cheaper (Schmeiser & Welch, 2006). Thus, automatic item generation is more cost efficient than traditional item writing (Kosh, Simpson, Bickel, Kellogg, & Sanford-Moore, 2019). Third, test security is improved because it is highly unlikely that two item sets are identical (Gierl & Lai, 2016; Wainer, 2002). Haladyna (2012, p. 13) concludes that item development is "the most expensive and most time-consuming aspect of test development", and automatic item generation tries to overcome this issue.

The main differences between automatic item generation and the approach in IRT and CTT is that the latter two techniques usually require all items to be tested before application and automatic item generation does not. Item calibration for each item is entirely eliminated in automatic item generation, as item difficulty is predicted by the item properties includeed rather than the item itself. This is because automatic item generation is based on an item model which states that distinct processes require a different amount of capability or capacity. If several processes are combined, the item difficulty is represented by the combined difficulty of each process. In order for this to be possible, several prerequisites need to be met. In general, three different approaches to automatic item generation have been identified: functional, model-based and automatic (Arendasy, Sommer, & Hergovich, 2007). In the functional approach, items are generated automatically without a cognitive model to serve as a foundation. In the model-based approach, item difficulty can be estimated based on the properties of the framework used.

Finally, in the automatic model, item difficulty can be predicted with a great deal of precision due to the use of a psychometric framework consisting of rules and categories. Typically, the test design matrix for automatic item generation is based on a theoretical model that has been empirically deduced or tested (e.g., Embretson & Kingston, 2018).

Different approaches to operationalization within automatic item generation are also possible. Irvine (2002) distinguishes between radicals and incidentals. Incidentals are irrelevant for predicting item properties and therefore represent only a superficial difference between items. Radicals, on the other hand, are integral for item difficulty and therefore substantially influence its prediction. Another approach is an item model (e.g., Bejar, 1996; Bejar et al., 2002; Bejar, 2002; LaDuca, Staples, Templeton, & Holzman, 1986), which consists of a stem, optional and auxiliary information (Gierl & Lai, 2016). The stem can be compared to a radical and the optional content to incidentals; ergo, the stem determines the item difficulty. Auxiliary information entails additional content like tables or graphs and can be assignet to either the stem or the optional information. Yet another approach is the cognitive design system approach (Embretson, 1998), which identified the cognitive abilities involved and links them to processing difficulty. The combination of cognitive processes involved determines the item difficulty; therefore, the item difficulty can be altered by changing the item properties involving a certain cognitive process. The advantage of this approach is that it not only determines item difficulty, but also identifies the source of cognitive capacity (Embretson, 1999). Item clones are essential for this model: “Item clones can be defined as generated items from a constant item form with some variable elements” (Embretson, 1999, p. 409). Altering these variable elements does not change the key component of the item; therefore, a person’s response to the old and new items should remain the same despite the presence of surface-level

differences. The original item is sometimes referred to as the parent item (Drasgow, Luecht, & Bennett, 2006).

Gierl and Lai (2016) describe the steps of automatic item generation in detail. First, the items' content is selected by an expert, taking into consideration that the items are to be generated automatically while still organizing and structuring content. Therefore, the second step is to select the specific content (e.g., concrete design attributes) for each item being created using a template, which functions as an item model (e.g., Bejar, 1996; Bejar et al., 2002; Bejar, 2002); it also also be referred to as a schema (Singley & Bennett, 2002), shell (Haladyna & Shindoll, 1989) or parent item (Drasgow et al., 2006), as mentioned earlier. This step potentially entails manipulating an item. In general, 1-layer models and n-layer models can be differentiated (Gierl & Lai, 2012b). In a 1-layer model, a relatively small number of attributes are changed to create more items. Normally, a parent item functions as a template to be manipulated (Gierl & Lai, 2016). The disadvantages of a 1-layer model are the limited number of possible items and the risk that the generated items may be too similar (isomorphic) to one another (Gierl & Lai, 2016). Taxonomies can help prevent too much similarity between items (Gierl, Zhou, & Alves, 2008). In the n-layer model, by contrast, more elements can be manipulated at the same time (Gierl & Lai, 2016). In the third step of the process, a computer generates the items at random by following the first two steps. Various software is available for this purpose (e.g., Higgins, Futagi, & Deane, 2005). One recommendation is to apply item precalibration to selected items (Sinharay & Johnson, 2012) to determine the item difficulty, since it is not realistic to test all items when there might be a thousand potential items (see next section).

In addition, similarity between items should be established. For instance, the word similarity index should be computed for word problems (Gierl & Lai, 2016). As in tests

constructed with IRT or CTT, a common problem in automatic item generation are the distractors in multiple choice items (Gierl et al., 2008; Gierl & Lai, 2016).

Psychometrics of automatic item generation. When constructing any test, the items must fit the model. For this purpose, Embretson (1999) recommends using LLTM (Fischer, 1973) to determine item difficulty parameters. Baghaei and Kubinger (2015) give quite detailed instructions on how to calibrate items within automatic item generation. The design or q-matrix/matrices, previously referred to as design matrix *w*, must be determined first. Q-matrices are used to reflect the item properties and therefore provide a solid estimation of a property's difficult level. If a model-based automatic item generation approach is chosen, these matrices tend to be based on theoretical models (Embretson & Kingston, 2018). The design or q-matrix depicts the processes or categories of each item. For example, a simple arithmetic task like 3 + 4 would have a 1 in the addition category and a 0 in the subtraction category in a q-matrix. In contrast, the task 7 + 2 – 5 would have a 1 in both of these categories. Moreover, these two problems vary in the number of terms involved. Hence, it would be reasonable to include a further category with the number of terms (see Table 2 for an example).

Table 2

Example q-matrix for arithmetic problems (without first column)

Problem	Addition	Subtraction	Three terms
3 + 4	1	0	0
7 + 2 – 5	1	1	1
6 – 5	0	1	0

The first column depicts the arithmetic problem and the three subsequent rows represent the design matrix. In the example outlined above, it is quite easy to determine the categories/processes. A key factor for a successful LLTM is the design of the weight matrix, because parameter estimations will fail otherwise and the parameters could be unreliable (Baker, 1993). Therefore, Baghaei and Kubinger (2015) propose trying different weight matrices and cross-validating them with a split dataset.

They suggest the following procedure for evaluating the fit of the LLTM (see also Figure 2; Baghaei & Kubinger, 2015): In a first step, the RM assumptions as well as the RM itself should be tested. In a following step, the LLTM is calculated. Before comparing the model's beta parameters, both beta parameters (the RMs and the LLTMs) should be normalized and correlated. Embretson (1999) reports a benchmark prediction level of at least $r = .70$ (e.g., Embretson, 1984, 1995, 1998; Embretson & Schneider, 1989; Whitely & Schneider, 1981). In a simulation study, Baghaei and Hohensinn (2017) come to a similar conclusion, and establish a benchmark of $r = .78$ for the item difficulty parameters of the RM and LLTM. Partial credit scoring seems to improve the prediction of item difficulty within automatic item generation (Diehl, 1998). Furthermore, Baghaei and Kubinger (2015) recommend calculating the difference in the -2 log-likelihood and calculating χ^2 under consideration of the degrees of freedom and an α-level of 5%, which is basically a test of deviances. To provide a better overview, a flow chart of the process is depicted in Figure 2.

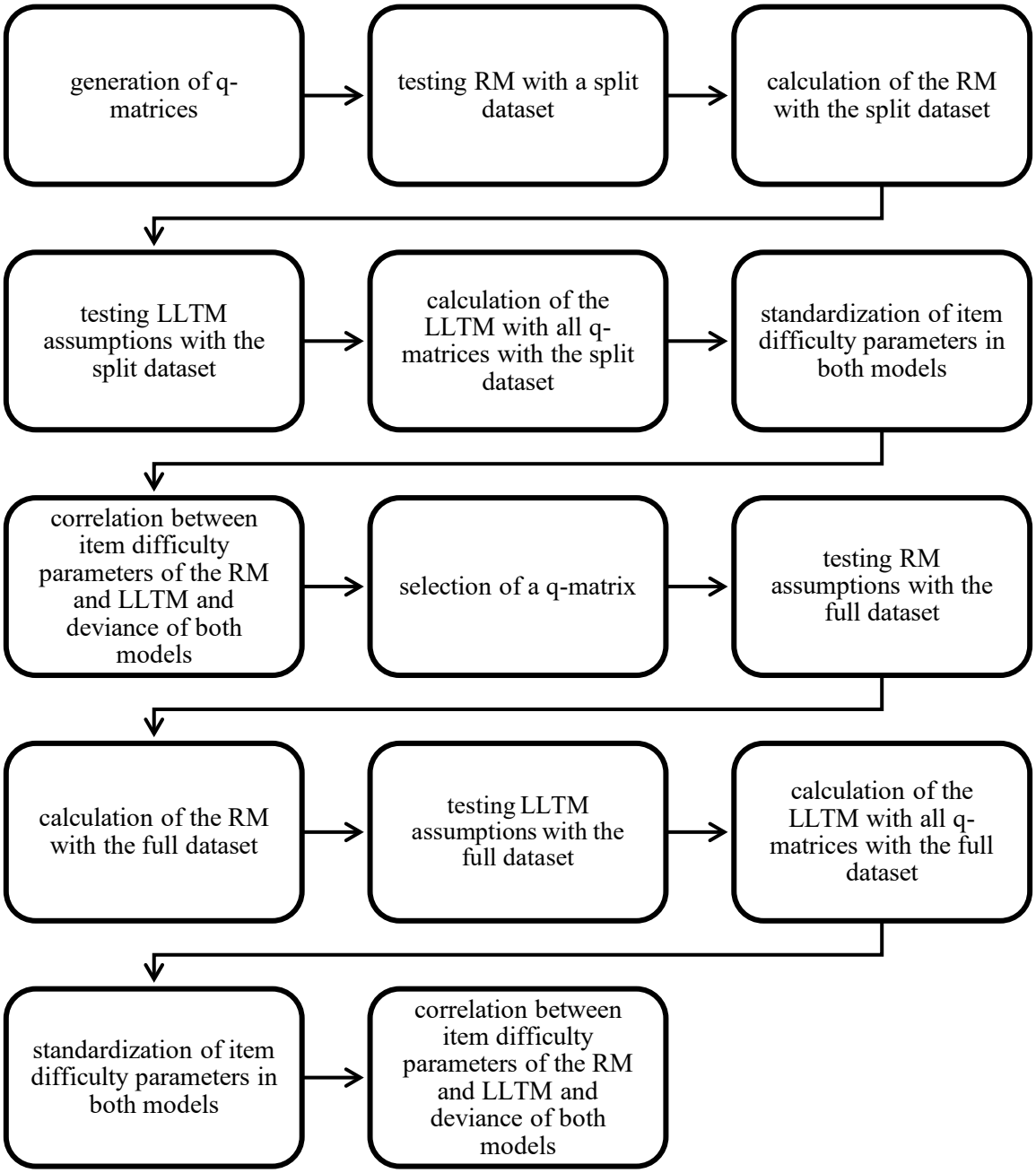


Figure 2. Flowchart of the LLTM evaluation process.

However, Baghaei and Hohensinn (2017) argue that this test seems too strict; instead, the correlation benchmark should be considered only if all other prerequisites are met. To establish this benchmark, Baghaei and Hohensinn (2017) propose three different

kinds of simulations. In the first step, a simulation with randomly generated q-matrices is conducted, with the proportion of 0's and 1's varied. Baghaei and Hohensinn (2017) suggest starting with a ratio of .30 1's and increasing the ratio up to 0.70, based on the observation that the proportion of 1's rarely exceeds 70%. Baghaei and Hohensinn (2017) themselves decided to increase from 30% to 70% work in 10 percentage-point steps, resulting in five steps. For each step, 1,000 matrices were produced, an LLTM calculated and the item difficulty parameter of the LLTM correlated with the item difficulty parameter of the RM. The minimum, maximum, median, mean and 95th percentile of all correlation coefficients generated for each 0-1 ratio was reported. Ideally, most correlations should be lower than the correlation coefficient obtained with the chosen q-matrix.

The second simulation suggested by Baghaei and Hohensinn (2017) permutates the rows of the original q-matrix. A total of 1,000 permutated q-matrices are randomly drawn, which are then used to calculate the LLTM. Subsequently, the item difficulty parameters of the LLTM are correlated with the item difficulty parameters of the RM and the same parameters reported (minimum, maximum, median, mean and 95th percentile of the correlation coefficients). The correlations should be higher than in the first simulation but lower than in the original q-matrix. Therefore, the first two simulations produce a lower benchmark (Baghaei & Hohensinn, 2017).

The third simulation aims to produce a perfectly designed q-matrix by reconstructing the item parameters. Again, 1,000 q-matrices are produced and the same coefficients reported as an upper benchmark.

These three simulations allow a benchmark to be established for the correlation between the item difficulty parameters of the RM and the item difficulty parameters obtained from the LLTM calculated with the designated design matrix.

Test Construction

A new test must also take some important test design aspects into consideration (e.g., Ziegler & Brunner, 2016). The American Educational Research Association (AERA), American Psychological Association (APA) and National Council on Measurement in Education (NCME) have developed joint guidelines for test development (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). Validity, reliability and fairness are among the most important principles that must be considered (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014) when constructing any test.

It is also important to consider the "ABC of test construction" (Ziegler, 2014, p. 239), which cover three aspects: definition of construct, intended use and targeted population. In addition, it should be considered who will employ the test later (Ziegler & Bensch, 2013). The purpose of the present project was to develop and evaluate a test tailored to the needs of the German Bundeswehr's OA, which follows a "select in" approach. For this reason, the test design is described in accordance with the requirements for this specific test.

Test Design

Choice of latent construct. The first step is to clearly define the purpose of the test. In the present case, the test should have high predictive validity for tests of on-site diagnostics. Ideally, the test should also be able to predict job performance. In addition, a construct must be chosen that can easily be measured online and is highly resistant to cheating.

As described above, the Bundeswehr selection process mainly focuses on intelligence (e.g., Krex, 2008). This is not surprising, as intelligence is considered the best singular predictor of job and training success (Ree et al., 1994; Schmidt & Hunter, 1981, 1998). However, most intelligence tests are long (e.g., Wechsler, 2008) and therefore presumably ill-suited for online use. It is also difficult to implement an intelligence test in a UIT that is robust against cheating.

WM, on the other hand, is very closely linked to various facets of intelligence, especially the measured aspects (see Chapter about Bundeswehr Recruiting and Personnel Selection): WMC can be considered an integral part of intelligence (Schneider & McGrew, 2012), can predict verbal intelligence (Süß et al., 2002), is closely related to arithmetic skill (Bayliss, Jarrold, Gunn, & Baddeley, 2003; Chen & Bailey, 2020; Fürst & Hitch, 2000; Lee, Ning, & Goh, 2013; Logie, Gilhooly, & Wynn, 1994), is associated with numerical intelligence (Süß et al., 2002), and predicts logical reasoning ability (Kyllonen & Christal, 1990; Oberauer, Süß, Wilhelm, & Wittmann, 2008) and fluid intelligence (Conway et al., 2002; Engle, Tuholski et al., 1999), the latter of which is relevant for progressive matrices (e.g., Lynn & Irwing, 2004). WM is further related to emotion regulation (Coifman et al., 2019; Schmeichel & Demaree, 2010) and emotion regulatory capacity (Coifman et al., 2019), which are also relevant for job performance (Newman, Joseph, & MacCann, 2010). In addition, WM correlates with abilities like task switching

(Baddeley, Chincotta, & Adlam, 2001; Vandierendonck, 2012; Vandierendonck, Liefooghe, & Verbruggen, 2010), self-regulation (Hofmann, Gschwendner, Friese, Wiers, & Schmitt, 2008), problem solving (Bühner, Kröner, & Ziegler, 2008; Hambrick & Engle, 2002, 2003; Wiley & Jarosz, 2012) and cognitive control (Kane, Conway, Hambrick et al., 2007), which seem to be needed in the work context as well.

Furthermore, WM seems highly relevant for complex military tasks (Nagler & Witzki, 2016) and predicts general situation awareness (Carretta, Perry Jr, & Ree, 1996), which is crucial for pilots. In addition, officer candidates must complete a university degree within the Bundeswehr. Since WM also predicts academic proficiency (Lee, Lee, Ang, & Stankov, 2009) and performance (Alloway & Alloway, 2010), this is a further argument in favor of measuring WM. These associations could also justify the use of WM for personnel selection in other armed forces (Irvine, 2014; Kemp & Jalbert, 2012; Wolgers, 2015).

Therefore, WM seems to be a suitable construct for personnel selection purposes in the military.

Task selection. As described above, OA poses a number of challenges that must be taken into account when designing a test (e.g., Schaper, 2009). For this reason, the test should be tailored to the needs of OA.

Dehn (2015) identifies four criteria for the measurement of WM: “integration of verbal and visual-spatial information; processing task-irrelevant, distracting information; ongoing inhibition, switching or updating [and] the conscious application of a strategy” (p.99). The test considered in the present dissertation needed to meet all four of these criteria.

In addition, the following requirements have to be met given that the test is to be conducted online: First, the test must be simple to understand, because only online instructions can be provided. Second, the test must be able to be completed on different electronic devices, like computers, smartphones and tablets, since their significance in pre-employment testing can be expected to increase (Illingworth, Morelli, Scott, & Boyd, 2015). Third, the test must be able to be completed using a computer mouse or a finger in the case of a smartphone or tablet. Fourth, cheating should be reduced to a minimum. Fifth, the test length must be a maximum of 15 minutes, including instructions. Otherwise, applicants could lose interest. Sixth, because applicants may be colorblind, the items can only be in black and white. Color can also have an influence on difficulty when it comes to visual WM (Morey, 2019). Seventh, test security needs to be maintained.

Considering the large number of Bundeswehr applicants (120,000 per year on average; 10,000 applicants for officer positions, Handelsblatt, 2019), creating a fixed set of items was not an option due to the risk of cheating in UIT (e.g., Arthur et al., 2009; Tippins et al., 2006). An immense pool of items (like in adaptive testing) could be the solution to that problem, with a new set of items drawn at random or based on ability level each time an applicant is tested. However, there are several problems to consider with this approach. First, the data (items) must be available and retrieved very quickly in order to avoid testing delays. Therefore, many servers need to run in the background to guarantee quick access. Second, many participants are needed to calibrate and evaluate all of the items. Third, the items must have similar or identical properties in order to ensure a fair testing procedure and stable quality criteria. Fourth, a starting theta and therefore a starting item set needs to be chosen. While many different approaches to choosing a initial theta estimate to start exist (see Magis, Yan, & Davier, 2017 for an overview), no satisfactory solution has been found, particularly one that considers the unique conditions in OA. For

example, if the first item were to always be the same, a solution could be quickly found online, distorting theta estimation when using a fixed number of items or stretching the length of the assessment (cf. Huang, 2018; Xu, Wang, & Shang, 2016).

Generating an algorithm that presents start items at random and is still able to estimate theta with sufficient precision would involve a great deal of effort and most likely a large number of items. This procedure would be inefficient and cost intensive because of the cost of item development (Rudner, 2010).

A second option would be automatic item generation, which avoids most of the problems associated with a large item pool like costs (Kosh et al., 2019). For this, a simple test was needed with items that could be generated based on a fixed set of relatively few rules. Conway et al. (2005) suggest measuring WMC with a WM span, as spans are valid and reliable. Although WM span tasks are frequently used (Conway et al., 2005), most spans are too easy to cheat on because they consist of numbers, words or letters (e.g., Case et al., 1982; Daneman & Carpenter, 1980; Oberauer et al., 2000; Turner & Engle, 1989; Unsworth, Redick, Heitz, Broadway, & Engle, 2009), which can be easily written down. Letter-number sequencing (Wechsler, 2008) was excluded for the same reason. In addition, traditional span tasks such as operation span were excluded because they are not suitable for an above-average ability sample (Draheim, Harrison, Embretson, & Engle, 2018). Hence, more complex visual and verbal stimuli are needed that are easy to display but hard to copy with pen and paper, and similar in a way that provokes interference.

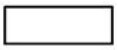
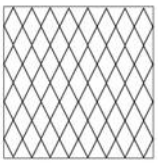
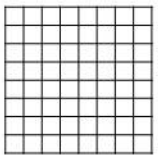
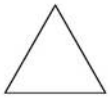
In addition, most WM spans are based on easy item generation rules and hence seemed perfect for automatic item generation. Therefore, an adapted complex span task seemed like a promising way to meet the requirements associated with measuring executive WM.

Operationalization. Thalmann and Oberauer (2017) comment on complex span tasks and possible interference as follows: “every model of WM must incorporate mechanisms that allow for domain-specific interference of cognitive and motor processing with both verbal and visuospatial memory” (Thalmann & Oberauer, 2017, p. 125). To provoke interference and generate many possible objects (instead of numbers, words or letters), a modular approach was chosen. This approach was chosen because only the same types of stimuli (in this case visual) provoke interference (Bae & Luck, 2019). Each object was made up of two parts: a background and a foreground layer. This approach was chosen to limit each object’s complexity, which can have an impact on recall (Oberauer & Eichenberger, 2013).

Each further layer was derived from the background to provoke interference. For this reason, the background had to consist of a geometric pattern. Because a verbal component was to be involved, the geometric shape making up the pattern needed to be easily recognizable and easy to name. As can be seen in Table 3, all shapes needed to have the same number of syllables (2) and approximately the same number of letters to avoid any bias due to word length or sound.

Table 3

Description of objects' backgrounds and foregrounds[8]

Geometric shape	Background	Figural layer	Verbal layer	Number of syllables	Number of letters
Rectangle			Rechteck	2	8
Rhombus			Raute	2	5
Square			Quadrat	2	7
Triangle			Dreieck	2	8

The advantage of this approach was that another processing component could be added to the test by assembling the objects in addition to just remembering the objects and repeating them in reverse order, as is common practice in simple span tasks (e.g., Baddeley, 2000; Oberauer et al., 2000). However, no distraction task was chosen as is usual for complex spans (e.g., Lewandowsky, Geiger, Morrell, & Oberauer, 2010; Oberauer et al., 2018), because that would immensely complicate automatic item generation. In addition, test security would be threatened, because it could become public knowledge that the distraction task is irrelevant for the task. Furthermore, instructing

[8] Since the test was conducted in German, the number of syllables and the number of letters refer to the German versions of the geometric shapes, which would be "Rechteck" (rectangle), "Raute" (rhombus), "Quadrat" (square) and "Dreieck" (triangle).

applicants on the task would become way more complex, and no person would be available to explain the task in case of questions or misunderstanding.

In the present task, participants see a series of objects and should recall them in reversed order. Since participants must assemble the previously shown objects by choosing two components (background and foreground), posing an additional challenge and including an additional processing component to provoke additional interference. Another advantage of this approach was that all possible responses could be displayed, as there would be a maximum of 12 options (4 backgrounds, 4 figural layers, 4 verbal layers). This solves the issue of distractors for multiple-choice items (Gierl et al., 2008; Gierl & Lai, 2016). Furthermore, such items such be easy to display on different media like tablets or smartphones. All in all, a total of 32 objects (four figural layers combined with four backgrounds and four verbal layers combined with four backgrounds, resulting in 16 combinations each) could be generated from the aforementioned backgrounds and foregrounds. An example object can be seen below (see Figure 3).

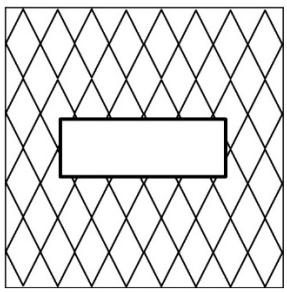


Figure 3. Example figural object.

Two different tests were created, because it was assumed that applicants for different jobs (privates/sergeants/non-commissioned officers and officer candidates) might have vastly different ability levels. Furthermore, Conway et al. (2005) argue that

individual measures of WMC are unable to measure WMC perfectly and hence suggest using multiple measures of WM to obtain a better result.

The figural WM test (WM-F) was the easier of the two. Only figural stimuli were combined to form a series that needed to be recalled. Therefore, only the objects with figural stimuli were used in the first test, resulting in a pool of 16 potential objects.

According to most literature (Alvarez & Cavanagh, 2004; Awh et al., 2007; Luck & Vogel, 1997), only four chunks can be held and processed in figural WM at a given time. However, applying a strategy and verbal encoding can enhance the span (e.g., Kliegl et al., 1987). Hence, the number of objects was limited to six, which was in accordance with the pretest (see section on the design of Study 1). Furthermore, chunking similar objects into clusters can boost performance (Son, Oh, Kang, & Chong, 2020).

To differentiate between different ability levels, items encompassing anywhere from one to six objects were built, resulting in six categories of items of increasing complexity (e.g., items in the first category consisted of only one object, items in the second category of two objects, etc.). This is comparable to Oberauer et al.'s (2000) approach for the DS backward. Because specific objects could not be repeated within one item (see next chapter), there was a pool of 16 possible items in the first category with only one object. In the second category, with items consisting of two objects, there were 240 potential items, and in the sixth and final category, with six objects per items, there were 5,765,760 potential items[9]. Given these examples, it should be obvious that generating the same item set twice within a short period is highly unlikely, even if only

[9] The number of items can be calculated by a simple formula given the condition that no object should be repeated: $\frac{n!}{(n-k)!}$ with *n* as the total number of available objects (16 in this case) and *k* as the number of objects that are drawn (for example, two in the second category), see Wagner (2006), for example. The exclamation mark refers to "factorial", where 3! equals 1*2*3 equals 6.

one potential item per category is selected. Even in this most basic case, 1,703,231,059,353,880,000,000,000 different item sets are possible.

For the first trial and the first test (WM-F), only two items from each category were chosen in order to keep the test within a reasonable length. Each item consisted of as many objects as the category dictated, which were presented quickly after one another. After all items were presented, participants had repeat the objects in reverse order. To do so, they had to select the foreground and corresponding background of each presented object. To reduce cheating, each object had to be entered within a certain amount of time; otherwise, the object was skipped and the next object had to be entered.

The choice of the exact response time was an important step, as it should be neither too short nor too long. Shepherdson, Oberauer, and Souza (2018) state: “Typically, as load increases, responses become slower, less accurate, or both” (Shepherdson et al., 2018, p. 286), which is a further indication that the time span should not be too short.

Since it takes a maximum of 300 ms to direct attention from one object to another (Hedge, Oberauer, & Leonards, 2015; Oberauer, 2003; Thigpen, Petro, Oschwald, Oberauer, & Keil, 2019) and participants needed some time to familiarize themselves with the response panel and to find the correct representations, it was assumed that a total of 10 seconds per object was sufficient.

A blue bar on the left corner indicated the remaining time. Furthermore, a red square indicated which object needed to be entered at the moment. In Figure 4, for example, the fourth object needed to be reproduced.

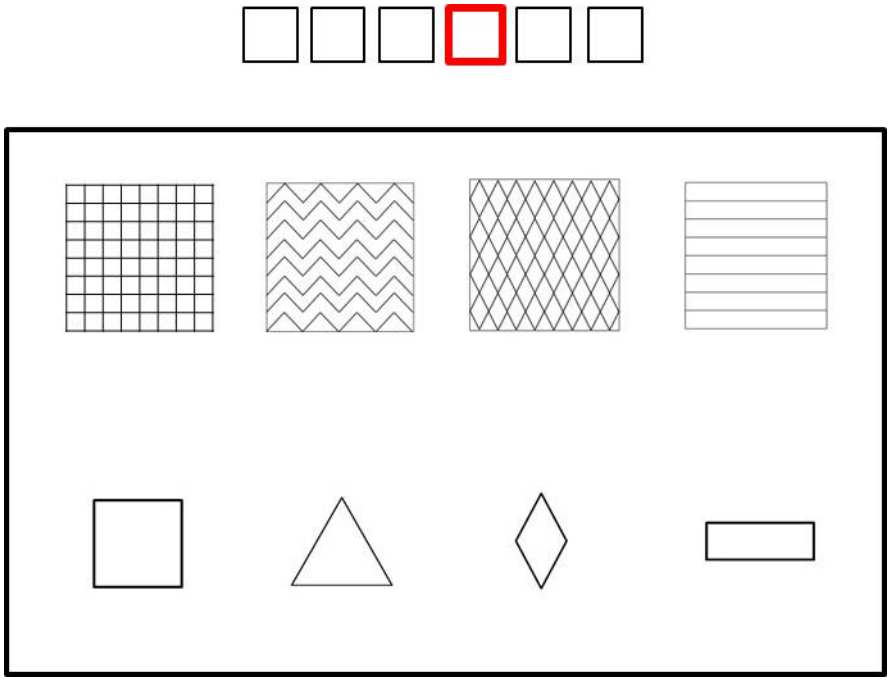


Figure 4. Example response panel for WM-F.

The second test, the verbal WM test (WM-V), mixed verbal and figural layers. As in the WM-F, item categories were defined according to the number of objects to be remembered and processed. However, in this test, not only figural but also verbal objects were included (see Figure 5 for an example).

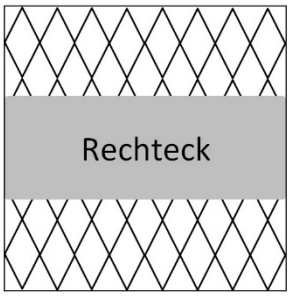


Figure 5. Example of a verbal object.

Verbal and figural objects were represented in alternation, with the verbal object always coming first. The objects had to be reproduced in reverse in this task as well. However, for verbal objects, the foreground (verbal) layer had to be reproduced by

selecting the corresponding figure. For example, if the verbal object in Figure 5 is presented, the figural object in Figure 3 must be built in the response panel (see Figure 6 depicted with arrows), making it more difficult to process the objects.

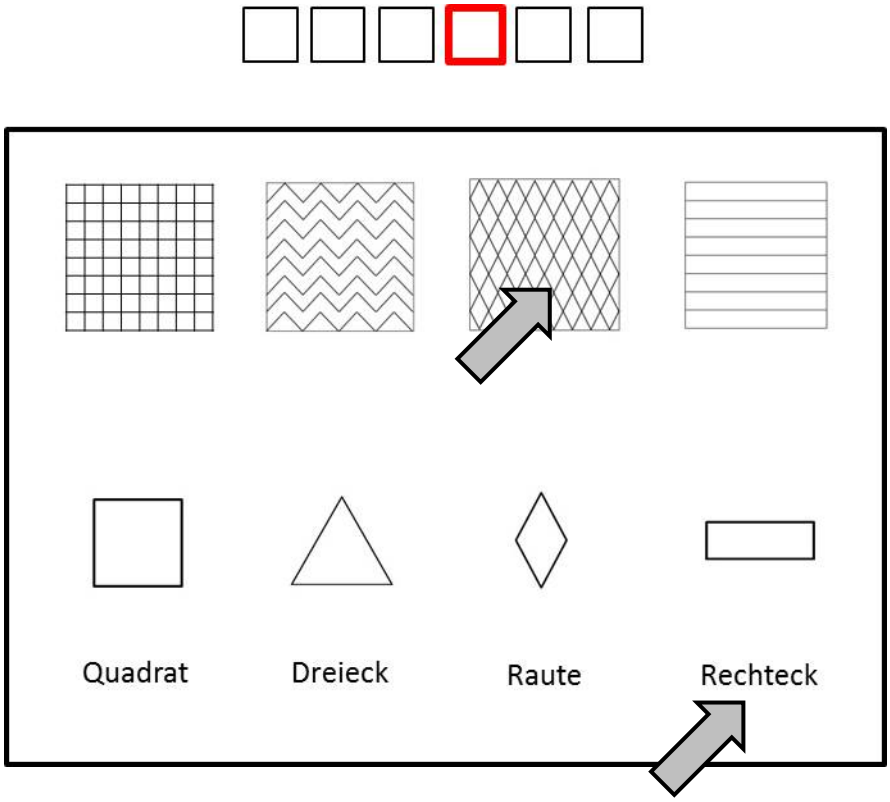


Figure 6. Answer screen for WM-V.

Specification of radicals and incidentals. In Irvine's (2002) understanding, the different objects would be incidentals, and test length the radical. Hence, item difficulty should only be determined by test length, not by the specific objects chosen. Therefore, the proposed tests are based on a 1-layer model (Gierl & Lai, 2012a).

Distractors. Since a modular approach was chosen, the answer screen could also be presented in such a way that test-takers had to put together the correct answer (see section on operationalization). Therefore, there is no need to create distractors that could have a significant impact on the item difficulty (e.g., D'Sa, Alharbi, & Visbal-Dionaldo, 2018).

Goals of the present project

The overall goal of the present project was to evaluate the proposed tests. The operationalized goal was to evaluate the proposed tests by testing whether automatic item generation is applicable (Study 1; Embretson, 1999) and whether the tests meet the requirements of objectivity, reliability and validity (Study 2; for an overview, see Bühner, 2011).In a first step, the model fit of the LPCM and LLTM needed to be determined (Study 1) in order to test whether automatic item generation was possible at all. Both models were needed because different scoring procedures could be applied. Either an all-or-nothing scoring approach could be selected, counting only items where all objects were chosen correctly, or a partial-credit scoring approach could be chosen, counting any and all correctly answered objects within each item (Conway et al., 2005). Therefore, the tests could fit the RM or the PCM and the corresponding linear versions.

In a second step, selected evidence was provided for the validity, reliability and scoring of the tests, with particular focus on their predictive power (Study 2).

Study 1

Methods

Sample. The full sample consisted of 330 participants. 82 were female and 248 were male. The average age was 22.65 (*SD* = 5.56). Most participants were currently attending or had graduated from an academic-track secondary school (German Gymnasium, *n* = 138) or middle-track secondary school (Realschule, *n* = 140). The remaining participants had graduated from a lower-track secondary school (Hauptschule, *n* = 52). The subsamples completing each item set are described in Table 4 to Table 6. Participants were randomly allocated to the item sets described below.

Table 4

Descriptive statistics of the subsamples for Item Sets 1 to 6

	I 1	I 2	I 3	I 4	I 5	I 6
N	62	62	57	47	55	52
n_{female}	18	16	17	10	10	11
n_{male}	44	46	40	37	45	41
Age	*M* = 23.64 *SD* = 6.05	*M* = 23.44 *SD* = 6.04	*M* = 21.83 *SD* = 6.04	*M* = 23.91 *SD* = 8.47	*M* = 21.58 *SD* = 4.38	*M* = 23.20 *SD* = 6.32

Note. I = item set.

Table 5

Descriptive statistics of the subsamples for Item Sets 7 to 12

	I 7	I 8	I 9	I 10	I 11	I 12
N	47	53	48	57	59	57
n_{female}	9	12	10	14	16	37
n_{male}	38	41	38	43	43	20
Age	M = 22.68 SD = 7.13	M = 23.31 SD = 5.83	M = 21.59 SD = 5.09	M = 23.49 SD = 7.39	M = 22.81 SD = 5.56	M = 23.03 SD = 6.05

Note. I = item set.

Table 6

Descriptive statistics of the subsamples for Item Sets 1 – 3, 4 – 6, 7 – 9 and 10 – 12

	I 1 – 3	I 4 – 6	I 7 – 9	I 10 – 12
N	181	154	148	173
n_{female}	51	31	31	50
n_{male}	130	123	117	123
Age	M = 23.01 SD = 6.10	M = 22.65 SD = 6.58	M = 22.57 SD = 6.08	M = 23.11 SD = 6.40

Note. I = item set.

Materials. Both WM tests (WM-F and WM-V) described in the previous chapter were applied. Due to the many possible items and thus broad range of potential item sets, only selected items were tested. The exact procedure is described in the next section.

Design. In a first step, the instructions were tested for comprehensibility and clarity due to their importance for OA. This took the form of structured interviews with $N = 10$ (8 male, 2 female; $M_{age} = 22.50$; $SD_{age} = 5.24$) participants. Participants were given the test instructions as well as standardized instructions telling them to read the instructions thoroughly. How long it took them to read the test instructions was measured (figural: $M_{time} = 298.76$ seconds, $SD_{time} = 116.25$ seconds; verbal: $M_{time} = 287.20$ seconds, $SD_{time} = 86.30$ seconds). After participants read the draft instructions, semi-structured cognitive interviews were conducted based on the following questions:

1. Did you have any problems understanding the instructions? If so, what did you not understand?
2. Could you explain what a task in this exercise looks like and how you have to complete it?
3. If you could change anything to make the instructions more understandable, what would it be?
4. Do you have any other comments concerning the test instructions?

After the first trial, the instructions for the other test were given and the procedure was repeated. To avoid any biases in understanding the first test, half the participants received the instructions for the figural WM test first and the verbal WM test second, whereas the other half received the instructions for the verbal WM test first and the figural WM test second. After the study, small changes were made to improve comprehensibility. Subsequently, the instructions were revised a final time using the Delphi method in an expert rating with two experts. Both experts were psychologists with experiencee in writing test instructions (years of experience: $M = 1.75$, $SD = 1.06$). Afterwards, the instructions were tested again with $N = 5$ participants. No further changes were made.

A second prestudy with $N = 10$ participants (5 male, 5 female; $M_{age} = 31.50$; $SD_{age} = 8.55$) evaluated how many objects could be remembered and processed. The goal was to determine an upper limit. Therefore, all participants had a university degree, since WMC and educational attainment seem to be associated with one another (Alloway & Alloway, 2010). Furthermore, Alloway and Alloway (2013) found that WMC is highest in 30 year olds.

The final version of the test instructions were applied. The maximum number of remembered objects was five. Since visual WMC seems to decline with age (Zhang, Shen, Tang, Zhao, & Gao, 2013), it was assumed that a maximum of six objects per item would be sufficient to differentiate between persons.

It was also tested how long each stimulus presentation should last. In similar tests, stimuli were presented for one second (Oberauer et al., 2000; Süß et al., 2002), but participants remarked that this duration was too short. In other WM tests that were somewhat less similar, presentation times of 3 seconds were chosen (Daneman & Carpenter, 1980; Turner & Engle, 1989). This was the reason for testing two different presentation durations (one and three seconds). An interstimulus interval (ISI) of 250 ms was selected based on similar tests, which had ISIs between 200 ms (Oberauer et al., 2000) and 300 ms (Salthouse & Mitchell, 1989). This also roughly matches the time it takes to shift one's attentional focus from one item in WM to another (Hedge et al., 2015; Oberauer, 2003; Thigpen et al., 2019).

The aim of the first study was to test the fit of the LLTM and LPCM with as many items as possible. Therefore, a balanced incomplete block design (BIBD) was applied (Frey et al., 2009). The rules for BIBD are:

1. Every cluster (t) occurs at most once in a booklet (b).

2. Every cluster appears equally often (r) across all booklets.

3. Every booklet is of identical length, containing the same number of clusters (k).

4. Every pair of clusters occurs together in booklets with equal frequency (λ).

(Frey et al., 2009, p. 45)

Because there were two different tests (figural and verbal), there were two BIBDs.

The clusters and corresponding items are provided in Table 7. Clusters A to C contain items from the figural WM test and Clusters D to F contain items from the verbal WM test. The category refers to the number of objects per item. Items in Category 1 consist of only one object, items in Category 2 of two objects, and so on.

Table 7

Clusters and corresponding items

Cluster	C 1	C 2	C 3	C 4	C 5	C 6
A	F101	F102	F103	F104	F105	F106
B	F201	F202	F203	F204	F205	F206
C	F301	F302	F303	F304	F305	F306
D	V101	V102	V103	V104	V105	V106
E	V201	V202	V203	V204	V205	V206
F	V301	V302	V303	V304	V305	V306

Note. C = category.

Each item set consisted of two clusters, analagous to the BIBD. The booklet design per test was balanced ($t = 3$, $b = 3$, $r = 2$, $\lambda = 1$). This resulted in three booklets per test with two different presentation times each, resulting in 12 different test designs total. The objects making up each cluster were drawn at random, but slightly modified according to the following rules to avoid biases in item sets, such as primacy or recency effects (Murdock, 1962), and to avoid chunking (Chase & Simon, 1973; Ein-Dor, 1971):

1. No items contain identical objects.
2. Maximum one-third of the sequence making up each item (sequence of objects) may be the same as the two following items.
3. The first and last objects of the two following items must be different.
4. Each object occurs at least once in a cluster.

Each item set consisted of two clusters. The items were combined while taking their category into consideration. For example, the order of items in Item Set 1 were F101, F201, F102, F202, F103 and so on (see Table 8). The number of items in each category was adapted from the DS backward task by Oberauer et al. (2000).

Table 8

Booklet design with corresponding clusters

Item set	Presentation time in s	Test	First cluster	Second cluster
Item Set 1	1	Figural	A	B
Item Set 2	1	Figural	C	A
Item Set 3	1	Figural	B	C
Item Set 4	3	Figural	A	B
Item Set 5	3	Figural	C	A
Item Set 6	3	Figural	B	C
Item Set 7	1	Verbal	D	E
Item Set 8	1	Verbal	E	F
Item Set 9	1	Verbal	F	D
Item Set 10	3	Verbal	D	E
Item Set 11	3	Verbal	E	F
Item Set 12	3	Verbal	F	D

Each participant completed both the figural and the verbal test. In order to avoid biases resulting from always presenting the same two item sets, each of the item sets within each test was combined with all possible item sets from the other tests, resulting in 36 different dyads (6 figural item sets * 6 verbal item sets), as can be seen in Table 9. At least 40 participants per item set were needed for a stable LLTM (MacDonald, 2014)[10], resulting in a sample size of 300 participants (50 * 6 item sets per test). However, in order to ensure that an approximately equal number of participants completed each trial, 324 participants plus 10 in case of participant dropout were recruited (300 participants minimum plus the number should be divisible by the number of trials (36)).

Table 9

Possible combinations of item sets

No	First trial	Second trial
1	Item Set 1	Item Set 10
2	Item Set 1	Item Set 11
3	Item Set 1	Item Set 12
4	Item Set 2	Item Set 10
5	Item Set 2	Item Set 11
6	Item Set 2	Item Set 12
7	Item Set 3	Item Set 10
8	Item Set 3	Item Set 11
9	Item Set 3	Item Set 12
10	Item Set 4	Item Set 7
11	Item Set 4	Item Set 8
12	Item Set 4	Item Set 9

continued

[10] Baker (1993) even suggests that a small sample size has little impact on the estimation of the parameters in the case of a dense design matrix.

continued

No	First trial	Second trial
13	Item Set 5	Item Set 7
14	Item Set 5	Item Set 8
15	Item Set 5	Item Set 9
16	Item Set 6	Item Set 7
17	Item Set 6	Item Set 8
18	Item Set 6	Item Set 9
19	Item Set 10	Item Set 1
20	Item Set 11	Item Set 1
21	Item Set 12	Item Set 1
22	Item Set 10	Item Set 2
23	Item Set 11	Item Set 2
24	Item Set 12	Item Set 2
25	Item Set 10	Item Set 3
26	Item Set 11	Item Set 3
27	Item Set 12	Item Set 3
28	Item Set 7	Item Set 4
29	Item Set 8	Item Set 4
30	Item Set 9	Item Set 4
31	Item Set 7	Item Set 5
32	Item Set 8	Item Set 5
33	Item Set 9	Item Set 5
34	Item Set 7	Item Set 6
35	Item Set 8	Item Set 6
36	Item Set 9	Item Set 6

Procedure. All participants had applied for a military career, ranging from private to non-commissioned officer. Data was gathered in seven participating Bundeswehr Career Centers (CC; Berlin, Dusseldorf, Erfurt, Hannover, Mainz, Stuttgart, Wilhelmshaven). Usually, applicants for positions from private to non-commissioned officers are tested in CCs with a similar personnel selection procedure.

In the CC, participants were asked whether they would be willing to participate in the study. After agreement, they received an informed consent form stating that participation was voluntary, explaining which data would be saved and processed and assuring that participation would have no effect on the hiring decision. Participants first completed all the psychological tests necessary to evaluate their job application on the computer. Subsequently, participants worked on the figural and verbal tests in accordance with the study design (see Table 9). Participants completed tests between 8 a.m. and 3 p.m., depending on their interview slot at the CC.

Statistical analysis. The statistical analyses in Study 1 had five aims. The first aim was to evaluate the q-matrices in order to choose one q-matrix design for later calculations. The second aim was to test the assumptions of the respective models. The third and fourth aims were to calculate and compare the respective models. The fifth aim was to describe how the subsamples match the calculated models across the complete dataset.

Evaluation of design matrices. First, the q-matrices for the LLTM were evaluated. The q-matrix with the best fit (highest correlation between item difficulty parameters of the RM and item difficulty parameters of the LLTM) was applied to the other datasets in order to validate the q-matrix (see Baghaei & Kubinger, 2015).

According to Baker (1993), misspecifications of the q-matrix have a greater impact on the estimation of parameters in a sparse q-matrix than in a dense q-matrix. Testing and cross-validating multiple q-matrices is in line with the approach by Baghaei and Kubinger (2015).

Two kinds of q-matrices are plausible, given that each dimension of a q-matrix represents a cognitive operation or category (Fischer, 1973; Sonnleitner, 2008). Since only two operations were applied to each item, namely repeating the items in reverse order and constructing the item, this approach does not seem advantageous. Instead, the categories should represent the number of objects to be remembered, since this is obviously the item property that determines its difficulty. As can be seen in Table 10 and Table 11, the sparse matrix contained 16.7% of 1's and the dense q-matrix contained 62.1% of 1's.

Table 10

Sparse q-matrix

Item	Number of objects	C 1	C 2	C 3	C 4	C 5	C 6
1	1	1	0	0	0	0	0
2	1	1	0	0	0	0	0
3	2	0	1	0	0	0	0
4	2	0	1	0	0	0	0
5	3	0	0	1	0	0	0
6	3	0	0	1	0	0	0
7	4	0	0	0	1	0	0
8	4	0	0	0	1	0	0
9	5	0	0	0	0	1	0
10	5	0	0	0	0	1	0
11	6	0	0	0	0	0	1
12	6	0	0	0	0	0	1

Note. C 1 = category contains only items with one object; C 2 = category contains only items with two objects; C 3 = category contains only items with three objects; C 4 = category contains only items with four objects; C 5 = category contains only items with five objects; C 6 = category contains only items with six objects.

Table 11

Dense q-matrix

Item	Number of objects	C 1	C 2	C 3	C 4	C 5
1	1	0	0	0	0	0
2	1	0	0	0	0	0
3	2	1	0	0	0	0
4	2	1	0	0	0	0
5	3	1	1	0	0	0
6	3	1	1	0	0	0
7	4	1	1	1	0	0
8	4	1	1	1	0	0
9	5	1	1	1	1	0
10	5	1	1	1	1	0
11	6	1	1	1	1	1
12	6	1	1	1	1	1

Note. C 1 = category contains items with at least two objects; C 2 = category contains items with at least three objects; C 3 = category contains items with at least four objects; C 4 category contains items with at least five objects; C 5 = category contains items with at least six objects.

To evaluate the q-matrices, the correlation between the beta parameter estimations from the RM and LLTM were calculated for both matrices. The same procedure was applied to the estimated thetas. Furthermore, the difference in -2 log likelihood was calculated. All results can be seen in Table 12. Model fits were the same for the dense and sparse q-matrices, although a high correlation of thetas can be expected in this kind of model. Therefore, Baker's (1993) recommendation to use the dense q-matrix was followed.

Table 12

Cross validation of the dense and sparse q-matrices across all item sets

Item set	Q-matrix	Correlation of beta parameters	Correlation of thetas	Difference in -2 log likelihoods
1	sparse	.95***	1.00***	11.23
	dense	.95***	1.00***	11.23
2	sparse	.94***	1.00***	39.75
	dense	.94***	1.00***	39.75
3	sparse	.73	1.00***	16.41
	dense	.73	1.00***	16.41
4	sparse	.97***	1.00***	28.93
	dense	.97***	1.00***	28.93
5	sparse	.97***	1.00***	33.58
	dense	.97***	1.00***	33.58
6	sparse	.98***	1.00***	18.46
	dense	.98***	1.00***	18.46
7	sparse	-	-	-
	dense	-	-	-
8	sparse	.98***	1.00***	3.43
	dense	.98***	1.00***	3.43
9	sparse	.99***	1.00***	2.59
	dense	.99***	1.00***	2.59
10	sparse	.94***	1.00***	15.74
	dense	.94***	1.00***	15.74
11	sparse	. 99***	1.00***	2.79
	dense	.99***	1.00***	2.79
12	sparse	.97***	1.00***	11.09
	dense	.97***	1.00***	11.09

Note. *** = $p < .001$.

Testing the assumptions. Unidimensionality was tested via item factor analysis ([IFA], Bock, Gibbons, & Muraki, 1988). The one-factor model was compared to the two-factor model using a likelihood ratio.

To test subgroup invariance and item homogeneity, a LRT (Andersen, 1973) and a Martin-Löf test (Christensen, Bjorner, Kreiner, & Petersen, 2002; Martin-Löf, 1973) were conducted. The LRT (specifically χ^2/df) seems to be a good measure of fit (Baghaei, Yanagida, & Heene, 2017) regardless of the sample size. Therefore, this fit measure is reported for the RM. The Martin-Löf test is recommended as well (Koller et al., 2012), even though it needs a large sample size to work properly (Verguts & Boeck, 2000). It is reported nonetheless for completeness' sake.

Although Verguts and Boeck (2000) recommend bootstrapping procedures for small samples (Davier, 1997), this method has not proven to be advantageous (Heene, Draxler, Ziegler, & Bühner, 2011). For this reason, bootstrapping was not used in this analysis.

In case of doubt (i.e., items had to be removed because of missing response patterns within subgroups), the T_{11}-statistic (Ponocny, 2001) was applied to the RM, because it only works with binary data. Furthermore, the T_{11}-test is most powerful for detecting violations in parallel item characteristic curves (Debelak, 2018).

Model calculation. For each item set, item parameters were calculated using the RM (Rasch, 1980), LLTM (Fischer, 1973; see also Fischer, 2005; Kubinger, 2009), PCM (Masters, 1982) and LPCM (Fischer & Ponocny, 1994; Fischer & Ponocny-Seliger, 1998).

Model comparison. To compare the models, item parameters from the RM and LLTM and from the PCM and LPCM were correlated with one another. This procedure is recommended by Baghaei and Kubinger (2015) for evaluating LLTMs. In this case, the same procedure was applied to evaluate the LPCMs. However, the -2 log likelihoods were not compared, since this approach seems to be outdated and does not lead to correct results, failing to appropriately consider the LLTM almost every time (Baghaei & Hohensinn, 2017). Instead, a benchmark for the correlation with the LLTM was used, as recommended by Baghaei and Hohensinn (2017).

For the LPCM, random q-matrices were generated at random with a replication rate of r = 1,000 and a proportion of 1's to 0's between 20% and 70%. The former was chosen because the q-matrix for an LPCM with three items in each category and six categories contains 20% 1's. The upper limit was selected in accordance with Baghaei and Hohensinn's (2017) recommendations. 5% steps were considered within this range in order to obtain a more precise result. In addition, a simulation with permutation was conducted. A q-matrix with the best possible fit was evaluated, and the beta parameter was correlated with the beta parameter of the PCM to obtain a high benchmark. The minimum, maximum and average correlation of the PCM parameters with the LPCM parameters were computed as benchmarks for the correlation between the LPCM and PCM. In case of imputed data, average beta parameters were calculated and correlated.

Furthermore, the 95th percentile was computed. The eRm package using conditional likelihood (Mair, Hatzinger, Maier, Rusch, & Debelak, 2019) was used to estimate the test parameters and an adapted version of the simulations by Baghaei and Hohensinn (2017) was used to fit the PCM and LPCM, using the eRm package (Mair et al., 2019) instead of the pcIRT package (Hohensinn, 2018).

According to Green and Smith (1987), the same persons should be deleted when calculating a LLTM as in a RM; thus, the same sample was used. The same procedure was applied to the LPCM and PCM. All items answered correctly by at least one person were included due to the relatively small sample size per item set. Consequently, it was possible for items to be included in the RM and LLTM comparison, but not the PCM and LPCM comparison, because they were not answered correctly by any participants.

Matching subsamples. As described earlier, multiple item sets were provided to different subsamples in a linked design. Hence, missing responses could be calculated to obtain item parameters for multiple item sets. This was achieved through multiple imputations with $k = 5$ (Li, Stuart, & Allison, 2015; Morris, White, & Royston, 2014) and predictive mean matching (pmm) to predict the missing values.

Results

The results of the prerequisite tests for the figural WM test (Item Set 1 - 6) are shown in Table 13.

Table 13

Results of tests for violations of assumptions

Item set	Test model	LRT	Martin-Löf	IFA[11]	χ^2/df
1	LLTM	.23	.95	.41	1.43
	LPCM	.98	.99	.22	
2	LLTM	.88	.73	.24	0.29
	LPCM	.84	.99	.21	
3	LLTM	X	X	.98	X
	LPCM	X	.96	.54	
1-3	LLTM	.39	.93	X	X
	LPCM	.72	.83	X	
4	LLTM	X	1.00	.08	X
	LPCM	X	1.00	.02	
5	LLTM	.48	.99	.18	0.89
	LPCM	.12	1.00	.20	
6	LLTM	.37	1.00	.26	1.06
	LPCM	.50	1.00	.01	
1-6	LLTM	X	X	X	X
	LPCM	.37	1.0	X	

Note. LRT = *p*-value of LRT; Martin-Löf = *p*-value of Martin-Löf test; IFA = *p*-value of IFA; χ^2/df = χ^2/df of LRT; X = value could not be obtained.

[11] Please note that all IFAs produce a Heywood case.

The correlations of the beta parameters from the RM or PCM with those from the LLTM, or alternatively LPCM for the figural WM test, are shown in Table 14.

Table 14

Overview of correlations for Item Sets 1 - 6

Test model	I1	I2	I3	I1-3	I4	I5	I6	I4-6
LLTM	.95***	.94***	.73	.95***	.97***	.97***	.98***	.96***
LPCM	.78***	.73***	.80***	.81***	.89***	.82***	.88***	.88***

Note. I = item set; *** = $p < .001$.

The results of the prerequisite tests for the verbal WM test (Item Sets 1 - 6) are shown in Table 15.

Table 15

Results of tests for violations of assumptions

Item set	Test model	LRT	Martin-Löf	IFA	χ^2/df
8	LLTM	X	.29	.23	X
	LPCM	.67	X	.19	
9	LLTM	.63	.10	.49	0.46
	LPCM	.74	.99	.05	
7-9	LLTM	.74	X	X	X
	LPCM	X	.16	X	
10	LLTM	.34	.72	.48	1.12
	LPCM	.25	.46	.91	
11	LLTM	X	.85	.30	X
	LPCM	.85	.18	.03	
12	LLTM	.82	.93	.02	0.38
	LPCM	.96	1.00	.40	
10-12	LLTM	.57	.04	X	X
	LPCM	.07	.53	X	

Note. LRT = *p*-value of LRT; Martin-Löf = *p*-value of Martin-Löf test; IFA = *p*-value of IFA; χ^2/df = χ^2/df of LRT; X = value could not be obtained.

The correlations of the beta parameters from the RM or PCM with those from the LLTM, or alternatively LPCM for the verbal WM test, are shown in Table 16.

Table 16

Overview of correlations of Item Sets 8 - 12

Test model	I8	I9	I7-9	I10	I11	I12	I10-12
LLTM	.98***	.99***	.95***	94***	.99***	.97***	.96***
LPCM	.73**	.72**	.79***	.82***	.88***	.91***	.86***

Note. I = item set; *** = $p < .001$; ** = $p < .01$.

Discussion

The overall aim of the first study was to determine whether automatic item generation would be possible with the presented test. The discussion is divided into several sections addressing different aspects of the results.

Assumptions. If the assumptions of the RM and the LLTM are not met, the informative value of the results is potentially low (e.g., Fischer, 1995; Wang & Wilson, 2005; Yen, 1993). Therefore, several tests were conducted to evaluate those assumptions.

A popular test is the LRT. In the present study, it was never significant. The LRT has been found to be particularly sensitive in detecting non-parallel item characteristic curves (Debelak, 2018), and the relatively small sample size should not be problematic either (Koller, Maier, & Hatzinger, 2015). However, a necessary condition for computing the LRT is that each response pattern exists at least once. Therefore, sometimes items needed to be excluded due to missing response patterns (see the Appendix for more thorough information on which items were excluded). In such cases, the test cannot provide information about all items, and the result may be biased. A common approach is to compute multiple LRT and adapt the *p*-value level accordingly with a Bonferroni correction. This was not possible either because the sample size was too small, meaning that even more items would have to be removed. However, this may only cause slight deviations in the LRT (Alexandrowicz & Draxler, 2016). In the end, all LRT with the imputed datasets were insignificant, indicating no violations of assumptions.

Baghaei et al. (2017) argue that the best fit index is Andersen's χ^2/df measure. In the present case, the RM fit was always under their suggested cut-off values. However, as

with the LRT, some items were removed because there were not sufficient response patterns.

A Martin-Löf test was also performed. This test was never significant either, with the exception of one imputed dataset (Item Set 10-12). However, it should be borne in mind that this test actually works better for larger samples, meaning that the power is low (Verguts & Boeck, 2000).

The IFA was also significant in some cases. Particularly notable that this test was always significant in the imputed datasets. For this reason, an additional simulation study was conducted to see whether this was due to the data imputation or the data structure (see Appendix). In general, the IFA seems to have problems with the present data structure (see Appendix, Study C). Since the simulation study was able to replicate this problem, the significant values should not be unduly considered. A further analysis revealed that the present data structure produces a Heywood case (Bock et al., 1988), meaning that the results of the IFA are not reliable.

Since the tests were mostly not significant, unidimensionality seems to be given and there is evidence for the validity of the test score interpretation. Nevertheless, Mair (2018) states that one should not only rely on numbers and significance values, but also critically examine whether item homogeneity exists. The results leave little cause for concern regarding unidimensionality, because unidimensionality can be seen as a continuum (e.g., Reckase, 2009) and the appropriate question to ask would be “at what point on the continuum does multidimensionality threaten the interpretation of item and person estimates?” (Smith, 2002, p. 206) rather than rating unidimensionality on a dichotomous scale. Since all items were constructed according to the same pattern and do

not show any significant semantic discrepancies, the test construction procedure allows for the assumption the sufficient homogeneity has been achieved.

All in all, it can be assumed that the assumptions are met and the informative value of the results is correspondingly high.

LLTM and LPCM fit. Overall, the LLTM and LPCM fit the data. The benchmark for correlations between the RM and the LLTM always exceeded the minimum benchmark for LLTMs of .78 (Baghaei & Hohensinn, 2017). Baghaei and Hohensinn (2017) even state that "when the LLTM perfectly fits we expect the correlation between RM item parameters and LLTM reconstructed parameters to be greater than $r = .95$. This scenario of simulations sets an upper bound for the expected correlation. Note that such a high magnitude of correlation is rarely obtained in practice as empirical data never perfectly fit mathematical models" (p. 898). Hence, both tests for the LLTM exhibit excellent results overall regarding the estimation of beta parameters. Thus, the automatic item generation seems to be valid for the present test procedures when all-or-nothing scoring is assumed, as is the case for the LLTM and RM.

For comparing the PCMs and LPCMs, no corresponding benchmark was available from the literature. For this reason, in accordance with Baghaei and Hohensinn (2017), simulation studies were carried out on the basis of the available data to establish a corresponding benchmark (see the appendix for the exact results). The benchmark was set to .80 based on the correlation that was closest to the expected value. It is clear that only parts of the LPCM worked, namely those in which stimulus presentation lasted 3 s. When the stimulus presentation time was shorter, the response patterns do not seem to be accurate enough for partial-credit scoring and therefore the automatic item generation does

not work as well as it should. This leads to the next important point, the duration of the stimulus presentation.

Duration of stimulus presentation. A stimulus presentation time of 3 s seems to be superior. The correlation coefficients for the figural WM test with a 1 s presentation time range from .73 to .95 for the LLTM and .73 to .81 for the LPCM, while the coefficients for the same test with a 3 s presentation time range from .96 to .98 for the LLTM and from .82 to .89 for the LPCM. A similar picture emerges for the verbal WM test. Here, the correlation coefficients with a 1 s presentation duration range from .95 to .99 for the LLTM and from .72 to .79 for the LPCM, while those with a 3 s interval range from .94 and .99 for the LLTM and from .82 to .91 for the LPCM. The better fit of the correlation parameter is clearly related to the longer duration of stimulus presentation. This is in accordance with Oberauer and Eichenberger's (2013) findings that encoding time and number of stimuli to process play an important role in visual WMC. Li, Xiong, Theeuwes, and Wang (2020) also found that a prolonged encoding time promotes visual WM.

Limitations

Sample size. There is no exact rule of thumb for sample sizes in terms of the LLTM, except for MacDonald's (2014) findings that confidence intervals become stable from 40 participants onwards. Nevertheless, larger sample size per item set would have been more advantageous, but was not feasible for organizational reasons. For this reason, the sample size was calibrated based on the required minimum.

Although Baker (1993) states that sample size has little influence on the estimation of parameters in an LLTM, differences in beta parameters for the different item sets were

found in the present case. For this reason, an additional simulation study was conducted (see appendix), which showed that a larger sample size (ideally at least $N = 250$) is preferable for the LLTM, particularly if not all items were answered correctly by at least one participant and the number of overall items decreases. However, O'Neill, Gregg, and Peabody (2020) found that although item calibration becomes less precise with decreasing sample size in the RM, person ability estimates are barely effected.

Sample. Considering that the study was conducted during the Bundeswehr's ongoing personnel selection process and participants were recruited during that process, a number of factors have to be taken into account. Although the testing was carried out in a controlled environment, some participants may have nevertheless cheated on the study. This is particularly likely if participants, despite the announcement that the study was independent of the selection process, did not believe this statement and therefore wanted to perform well. Closely related to this question is whether participants really believed that partication was voluntary or considered the experiment to be a hidden test. This could also lead to inconsistencies in the response patterns.

Another point that it cannot be assumed with certainty is that all participants were equally motivated to participate in the study. Such differences could also lead to inconsistencies in response behavior.

Furthermore, participants were under some degree of pressure, because although the psychological testing procedures had been completed, they were still in the middle of an application process. This could also have had an impact on their performance.

For organizational reasons, officer candidates could not be tested in the first study, which probably led to a restricted range of abilities. Another weak point is that few women

participated in the study, although this corresponds to the Bundeswehr's overall applicant pool, since on average more men apply for military professions.

Test setting. A further critical point concerns the test environment in which the study was conducted. Because testing took place during normal recruitment operations, the sample perfectly corresponded to the target group, but the conditions could not be kept exactly the same for all subjects. Although the test administrators were thoroughly trained and received specific standardized instructions for conducting the study, the fact that the participants took the test with different administrators in different rooms may well have an influence. In addition, variables such as test duration varied greatly between morning and afternoon participants. Furthermore, different subjects applying for different positions with the Bundeswehr had been administered different tests before beginning the study and thus started the experiment in different circumstances and with different states of depletion. This may be problematic, since previous tasks may influence performance on complex span tasks (e.g., Healey, Hasher, & Danilova, 2011), even though WM variation over the day is very small (Gevins et al., 2012) and relatively stable over time in general (Xu et al., 2018). These environmental conditions make it difficult to determine the maximum ability of a particularly competent subject in the present study. An additional complicating factor for a clean diagnosis of WMC is that quality of visual WMC can vary during tasks (Fougnie, Suchow, & Alvarez, 2012).

However, these non-standardized conditions and corresponding results do not pose a serious problem, since subsequent real-world testing will also be carried out under varying conditions. If the test model works nevertheless, this is a good indicator that it can also work under real conditions.

Conclusion

Despite some limitations, the tests produce data that are in line with the test models' assumptions.

Overall, it can be assumed that both test procedures are suitable for automatic item generation. This is especially true with a presentation time of 3 s. Furthermore, all-or-nothing scoring seemed to be superior compared to partial-credit scoring. Therefore, the longer stimulus presentation time was selected for the following study. Special attention must be paid later on to the selection of the scoring technique, since qualitative discrepancies were already apparent in the first study. However, the final decision of whether to select all-or-nothing scoring or partial-credit scoring must be made in the next study based on an external criterion.

Since almost no participants succeeded in remembering six objects in the WM-V, such items can be excluded.

Based on the second simulation study, the next study should entail at least $N = 250$ participants for every test. This should be a sufficient sample size for detecting weak violations of the assumptions as well (Koller et al., 2015). The sample should also be extended to include officer candidates for two reasons: firstly, this would represent a wider range of abilities, and secondly, the relevant clientele must be tested in order to determine the predictive validity of the test procedures. Furthermore, it would be advantageous to increase the number of items in order to facilitate more differentiated diagnostic decisions based on persons' specific ability parameters.

Study 2

Introduction

I will first briefly shed light on the relevant background to Study 2. The goal was to make a final decision on a test model and provide evidence of the tests' validity and reliability. In order to improve the validity of the test scores, the number of items was increased to create more variance between subjects. Moreover, the stimuli presentation time was set to 3 s instead of 1 s.

Since the tests produced data in line with the assumptions of the LLTM and LPCM in Study 1, it can be assumed that this will be the case again in Study 2. However, it is expected that one of the tests will be better suited for OA with the respective target groups (namely officers and privates/sergeants/noncommissioned officers) based on its validity and reliability. Therefore, one aim of Study 2 was to decide whether to implement the WM-F or WM-V for the Bundeswehr's OA.

As previous described, there are different facets of WM. This seems to be reflected in the different results produced by different measures of WM. For example, complex spans (e.g., Oberauer et al., 2003) correlate only relatively weakly with n-back tasks (Kane, Conway, Miura et al., 2007; Oberauer, 2005a) or the DS backward (Hilbert, Nakagawa, Puci, Zech, & Bühner, 2015), even though all of these tests claim to measure WM. Oberauer et al. (2003) were able to assign WM tests to various functional factors, namely storage and processing, supervision/switching and coordination. Therefore, there seem to be multiple mechanisms at work in WM. Storage and transformation encompasses transforming information and storing them accordingly. Supervision tasks entail "selectively activating relevant representations and processes and inhibiting irrelevant ones" (Oberauer et al., 2000, p. 1019). Finally, coordination is understood as linking different objects to their distinct location or position (Oberauer et al., 2000). This entails

constructing an image out of its component parts (cf. Kosslyn, Reiser, Farah, & Fliegel, 1983), for example. Furthermore, as described above (Oberauer et al., 2003), two content factors can be assumed, namely spatial and non-spatial (verbal-numerical). WM tests can be clustered using this classification system to make differences and similarities within measures more visible. For example, the test procedures in the current project are classified in Table 17. In comparison, Oberauer et al. (2000) classifies the DS backward as storage and transformation. Categories that apply to the tests are marked with an X.

Table 17

Classification of WM tests

Test	Functional factors			Content factors	
	Storage and transformation	Supervision	Coordination	Spatial	Verbal-numerical
WM-F	X		X	X	
WM-V	X	X	X	X	X
DSB	X				X

Note. DSB = DS backward

As Oberauer et al. (2003) demonstrate evidence for the aforementioned categories by means of factor analysis, they could therefore be responsible for the different correlations among WM tests. Therefore, it can be assumed that both tests (WM-F and WM-V) correlate with the DS backward (evidence for convergent validity of the test scores). However, this correlation should not be as high as in a regular comparison of evidence for convergent validity (e.g., Bühner, 2011) since the tests do not cover the same functional factors.

These functional factors can influence the assocation between WM and higher-order cognition as well (Unsworth, Redick et al., 2009): complex span and n-back tasks account for different kinds of variance predicting Gf (Kane, Conway, Miura et al., 2007), for example. Therefore, the following benchmarks for correlations between WM and higher-order cognition are always average correlations for different kinds of WM measures.

In the past, various studies have been conducted comparing the correlations between performance on matrices tests and WM tests. An average correlation of $r = .30$ (Conway et al., 2002; Kane et al., 2004; Kane, Conway, Miura et al., 2007; Unsworth, Brewer, & Spillers, 2009; Unsworth & Engle, 2005; Wiley, Jarosz, Cushen, & Colflesh, 2011) has been found. Since the present tests were also intended to measure WM, they should correlate at a similar level with the Bundeswehr matrices test. In addition, WM tests have divergent correlations with verbal analogies, ranging from $r = .06$ to $r = .35$ (Unsworth, Brewer et al., 2009). The tests most similar to the WM-F and WM-V have correlations of $r = .28$ (item recognition with pictures) and $r = .10$ (picture source recognition; Unsworth, Brewer et al., 2009), respectively. For this reason, the presented tests should correlate approximately similarly with verbal analogies. Since the WM-V focuses more strongly on verbal content, it should have a higher correlation with the verbal analogies test than the WM-F.

Furthermore, arithmetic tests have shown a medium correlation with WM tests of about $r = .33$ on average (e.g., Friso-van den Bos, van der Ven, Kroesbergen, & van Luit, 2013; Peng, Namkung, Barnes, & Sun, 2016). Therefore, the present tests should exhibit an equally high correlation. Since the verbal and numerical facet of WM seems to be one content factor within WM (Oberauer et al., 2003), WM-V should correlate more strongly with the arithmetic test.

Since no have examined the composite scores currently used by the Bundeswehr, it must be assumed that, similarly to the matrices test, the verbal analogies and arithmetic tests correlate equally strongly with WM-F and WM-V. Since both the matrices test and verbal analogies test can be regarded as latent measures of fluid intelligence (e.g., Unsworth, Brewer et al., 2009), measures of which correlate at a level of $r = .30$ with WM tests (Unsworth, Brewer et al., 2009), a correlation of approximately $r = .30$ can be assumed. In addition, since WM-V better captures the verbal factor of WM, the correlation with WM-V should be higher than the correlation with WM-F.

When it comes to structural validity, the choice of a scoring model is important (Messick, 1995). The scoring model should optimally fit the measured construct. Therefore, the model fit and correlations with DS backward should be taken into account. In the end, only one scoring model should be chosen. Furthermore, complex spans like the tests developed in this project exhibit moderate to high internal consistency (e.g., Conway et al., 2002; Engle, Tuholski et al., 1999; Unsworth, Heitz, Schrock, & Engle, 2005). Therefore, the WM-F and WM-V should also exhibit internal consistency estimates at this level.

Although there is some evidence for gender differences in WM (e.g., Saylik, Raman, & Szameitat, 2018), such differences should not be relevant for the present study, since the effect is rather small for visual-spatial WM (Voyer, Voyer, & Saint-Aubin, 2017). Therefore, neither test should exhibit DIF (differential item functioning).

Methods

Sample. The full sample consisted of 621 participants, 142 female and 479 male. The average age was 23.54 (*SD* = 5.95). A majority of participants were attending or had graduated from an academic-track secondary school (Gymnasium, *n* = 321) or middle-track secondary school (Realschule, *n* = 216). The remaining participants had graduated from a lower-track secondary school (*n* = 82) or special needs school (*n* = 2). The two subsamples taking the two different tests are described in greater detail below.

The subsample for the figural WM test consisted of 316 participants (*n* = 242 male and *n* = 74 female). The average age was 23.99 (*SD* = 6.55). 164 participants had graduated from or were attending an academic-track secondary school, 104 participants had graduated from or were attending a middle-track secondary school, and 48 participants had graduated from a lower-track secondary school.

The subsample for the verbal WM test consisted of 305 participants (*n* = 237 male and *n* = 68 female). A majority of participants were attending or had graduated from an academic-track secondary school (*n* = 157), 112 participants were attending or had graduated from a middle-track secondary school, 34 participants had graduated from a lower-track secondary school and 2 participants had graduated from a special needs school.

Materials. The same WM tests were used as in Study 1. However, in this study, the WM-F contained six categories (with 1-6 objects per item) and four items per category, resulting in 24 items, and the WM-V contained five categories (with 1-5 objects per item) and four items per category, resulting in 20 items in total.

The average processing time was *M* = 10 min 11 s (*SD* = 1 min 24 s) for the figural test and *M* = 8 min 15 s (*SD* = 1 min 25 s) for the verbal test.

For validation purposes, a DS backward (e.g. Wechsler, 2008) was applied as well. Furthermore, three tests already being used in the Bundeswehr personnel selection process were considered: a verbal analogies, arithmetic and progressive matrices tests (see section "Bundeswehr Recruiting and Personnel Selection" for a more detailed description).

Design.

Procedure. As in Study 1, all participants were applicants for a military career, ranging from applicants aiming to be a private to officer candidates. Data was gathered in six participating CC (Berlin, Dusseldorf, Erfurt, Hannover, Stuttgart, Wilhelmshaven) and the AC (Assessmentcenter für Führungskräfte der Bundeswehr / Assessment Center for Bundeswehr Officers in Cologne). Officer candidates are tested at the AC. Different procedures had to be followed at the two types of testing sites due to differences in the application process.

Participants in the CC were asked whether they would be willing to participate in the study. After agreement, they received an informed consent form. It stated that participation was voluntary, which data would be stored and processed and that their participation would have no effect on the Bundeswehr's hiring decision. Participants first completed all the psychological tests necessary to evaluate their job application on the

computer. Subsequently, participants worked on the WM-F or the WM-V. Afterwards, a DS backward was conducted. Participants completed tests between 8 a.m. and 3 p.m., depending on their interview slot at the CC.

In the AC, a different procedure was applied. Applicants arrived on the day before the assessment and heard a presentation about the assessment procedure that would take place over the following days, as is standard procedure (Bundeswehr, 2014; see also section "Bundeswehr recruiting and personnel selection"). After the presentation, standardized recruitment for the study took place. Volunteers participated in the study immediately afterwards. Thus, all officer candidates took the test at around 5 p.m. from Monday to Wednesday.

Statistical analysis. To test the model fit, RM, LLTM, PCM and LPCM were estimated in the same way as in Study 1, although there were no multiple imputations and no validation of the design matrix. The same dense design matrix was used as in Study 1. The tests' structural validity was examined through Pearson correlations with the DS backward, thetas from the progressive matrices test, thetas from the verbal analogies test and thetas from the arithmetic test as well as an overall score. Pearson correlations were calculated with 2,000 bootstrapped samples (Field, Miles, & Field, 2012) to obtain bias-corrected and accelerated (BCa) 95% confidence intervals ([CI], e.g. Efron, 1987) using the boot package in R (Canty, 2002). The tests' validity was examined by conducting ROC analysis (see Fawcett, 2006 for an overview) using the pROC package (Robin et al., 2011) with the corresponding cut-off values for the overall test score in the application process. Because the goal was to choose the best applicants, the cut-off value was set to the best possible result in personnel selection.

A 95% confidence interval was calculated for the area under the ROC curve (AUC). In order to define a threshold value for diagnostics, Youden's index (Youden, 1950) was also calculated.

Split-half reliability scores were obtained by splitting each test into two scores along different split criteria, with groupings for the first and second items within a category, first and third items within a category, and first and fourth items within a category. Antagonistic scores were obtained in an inverse process, e.g. the second and third items in a category were compared to first item. Both scores were correlated and corrected with the Spearman-Brown formula.

Fairness was tested via DIF using gender as a split criterion and logistic regression (Choi, Gibbons, & Crane, 2016; Zumbo, 2007) with the lordif package (Choi et al., 2016).

Results

Model fit. The results of the prerequisite tests for both tests are shown in Table 18.

Table 18

Results of tests for violations of assumptions

Test	Test model	LRT	Martin-Löf	χ^2/df
WM-F	LLTM	.24	.12	1.20
	LPCM	.42	1.00	
WM-V	LLTM	.73	1.00	0.70
	LPCM	.02	1.00	

Note. LRT = *p*-value of LRT; Martin-Löf = *p*-value of Martin-Löf test; χ^2/df = χ^2/df of LRT.

The correlations between the beta parameters of the RM or PCM and the LLTM or LPCM for both tests are shown in Table 19.

Table 19

Overview of correlations of both tests

Test model	WM-F	WM-V
LLTM	.98***	.97***
LPCM	.86***	.88***

Note. I = item set; *** = $p < .001$.

Validity. Correlation coefficients between the proposed tests and the tests currently used in personnel selection can be seen in Table 20 and Table 21 (evidence for convergent and discriminant validity).

The correlations between the figural WM test and the tests selected to assess criterion validity can be found in Table 20. The second column (DS backward) depicts convergent validity, while the third to sixth columns depict discriminant validity.

Table 20

Correlations of figural WM test scores with the diagnostic assessment

WM-F	DS backward	Matrices	Verbal analogies	Arithmetic	Composite
LLTM	.38***	.25***	.13*	.27***	.27***
	CI (BCa)	CI (BCa)	CI (BCa)	CI (BCa)	CI (BCa)
	[.29, .46]	[.11, .36]	[-.02, .25]	[.14, .38]	[.11, .41]
	n = 315	*n* = 315	*n* = 312	*n* = 282	*n* = 282
LPCM	.47***	.28***	.12*	.29***	.29***
	CI (BCa)	CI (BCa)	CI (BCa)	CI (BCa)	CI (BCa)
	[.37; .55]	[.15, .40]	[-.03, .25]	[.16, .40]	[.13, .40]
	n = 315	*n* = 315	*n* = 312	*n* = 282	*n* = 282

Note: CI = confidence interval; BCa = bias-corrected and accelerated; *** = $p < .001$, * = $p < .05$.

The correlations between the figural WM test and the tests selected to assess criterion validity can be found in Table 21. The second column (DS backward) depicts convergent validity, while the third to sixth colums depict discriminant validity.

Table 21

Correlations of verbal WM test scores with the diagnostic assessment

WM-V	DS backward	Matrices	Verbal analogies	Arithmetic	Composite
LLTM	.25***	.29***	.28**	.35***	.38***
	CI (BCa)	CI (BCa)	CI (BCa)	CI (BCa)	CI (BCa)
	[.15, .35]	[.18, .39]	[.17, .38]	[.24, .43]	[.29, .47]
	n = 305	*n* = 268	*n* = 268	*n* = 263	*n* = 263
LPCM	.34***	.30***	.27**	.38***	.39***
	CI (BCa)	CI (BCa)	CI (BCa)	CI (BCa)	CI (BCa)
	[.23; .43]	[.18, .42]	[.16, .37]	[.28, .46]	[.29, .48]
	n = 305	*n* = 268	*n* = 268	*n* = 263	*n* = 263

Note: CI = confidence interval; BCa = bias-corrected and accelerated; *** = $p < .001$, ** = $p < .01$.

The results of the ROC analysis of both tests on the composite of the tests used in personnel selection can be found in Table 22, providing evidence of predictive validity. The corresponding ROC curves for the WM-F can be found in Figure 7 and Figure 8. Figure 9 and Figure 10 depict the ROC curves for the WM-V.

Table 22

Results of ROC Analysis for WM-F and WM-V

Test	Test model	AG	AUC	CI_{low}	CI_{high}	Youden	Spec	Sens
WM-F	LLTM	O	.80	.64	.90	8.5	.81	.70
		P	.64	.56	.73	6.5	.55	.71
	LPCM	O	.71	.53	.84	38.5	.76	.60
		P	.66	.58	.74	25.5	.35	.92
WM-V	LLTM	O	.73	.57	.83	3.5	.51	.85
		P	.69	.60	.77	3.5	.55	.72
	LPCM	O	.74	.61	.84	14.5	.55	.92
		P	.67	.60	.75	13.5	.56	.76

Note: AG = applications group; AUC = Area under the ROC Curve; CI = confidence interval; O = officer; P = privates/sergeants/non-commissioned officers; Youden = Youden's index; Sepc = specifity; Sens = sensitivity.

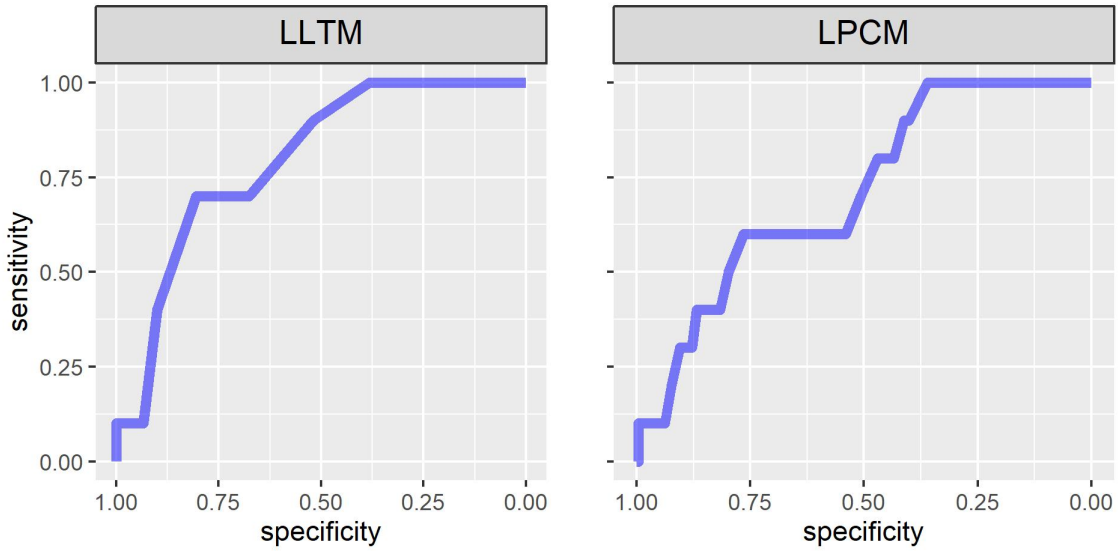


Figure 7. ROC curves for WM-F (officer candidates).

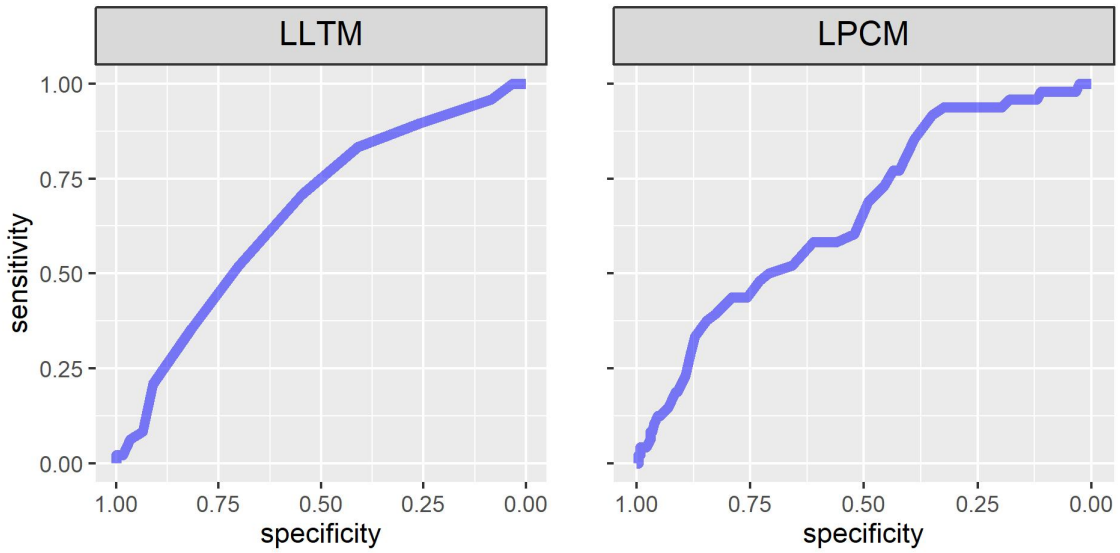


Figure 8. ROC curves for WM-F (privates/sergeants/non-commissioned officers).

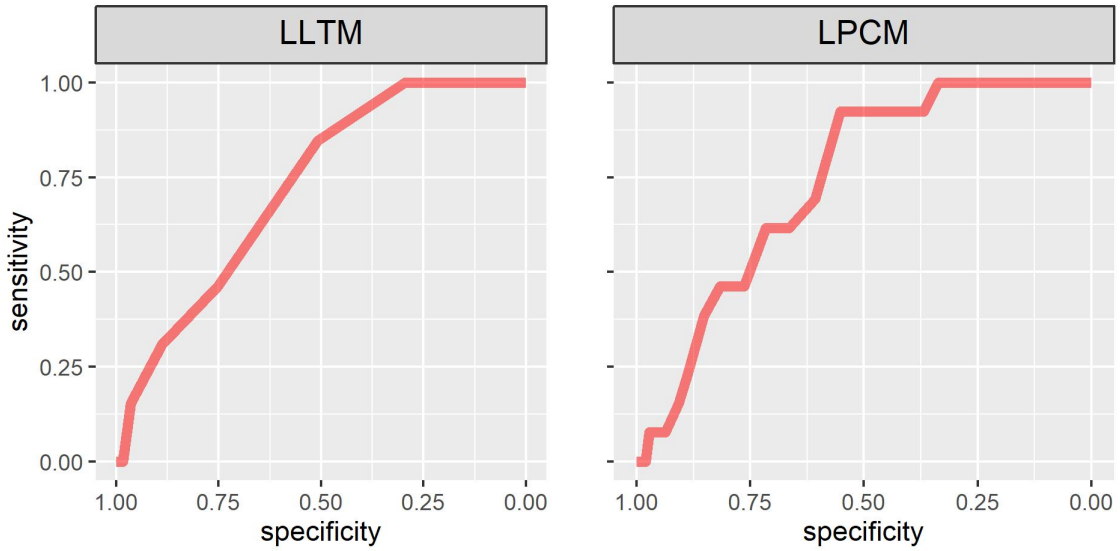


Figure 9. ROC curves for WM-V (officer candidates).

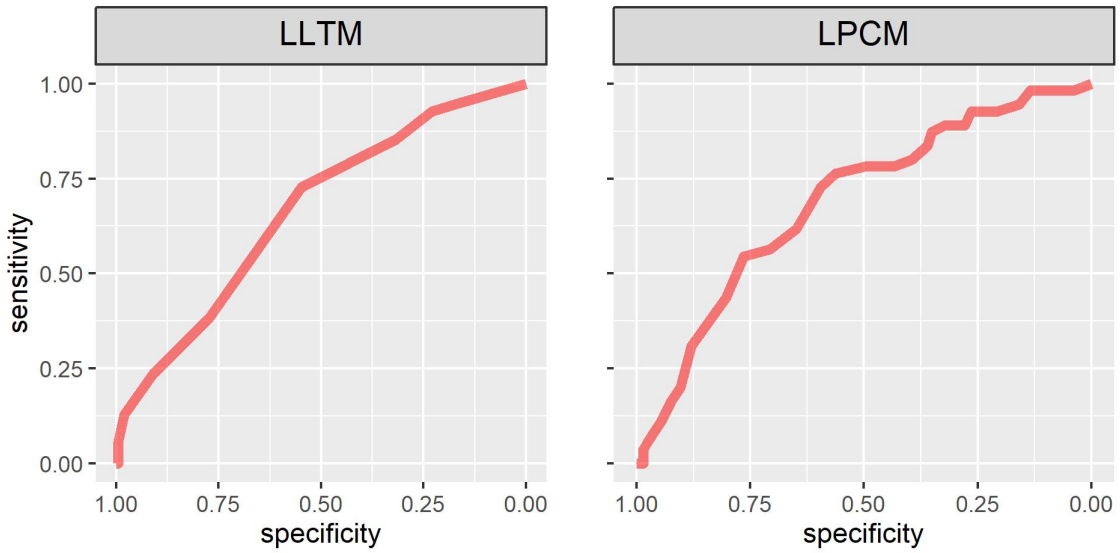


Figure 10. ROC curves for WM-V (privates/sergeants/non-commissioned officers).

Reliability. Split-half reliability was calculated and corrected with the Spearman-Brown formula. Results can be seen in Table 23.

Table 23

Split-half reliability correlations for WM-F and WM-V

Test	Test model	SC1	SC2	SC3
WM-F	LLTM	.88***	.89***	.91***
	LPCM	.74***	.74***	.75***
WM-V	LLTM	.90***	.90***	.89***
	LPCM	.73***	.80***	.77***

Note: SC = split criterion; *** = $p < .001$.

Fairness. Due to missing response patterns, DIF through logistic regression could only be calculated for the LPCM of WM-V. No items showed DIF ($p < .01$) by gender.

Discussion

Model fit. As in Study 1, most of the tests indicated no violations of the assumptions. Since Andersen's χ^2/df remained below the proposed limit in all cases (Baghaei et al., 2017) and the LRT and Martin-Löf tests were not significant (with one exception), the assumptions can be considered met.

As in Study 1, the LLTM fit the data better than the LPCM, suggesting that all-or-nothing scoring should be applied to both tests.

Thus, it has been repeatedly shown that automatic item generation works for the present test procedures. The LLTM, in particular, achieved a perfect fit based on Baghaei and Hohensinn's (2017) reference value.

Validity. When evaluating validity, several aspects can be considered. The following section first addresses convergent and discriminant validity, for which the results can be found in Table 20.

Convergent validity. The correlation between the test procedures and the DS backward meets and even exceeds the expected value (Hilbert et al., 2015). The weak correlation coefficient might be explained by discrepancies in which aspects of WM the measure covered, as previously described (see Oberauer et al., 2000). Since WM-F does not cover the supervision aspect of WM like the DS backward as well, it is reasonable that the corresponding correlation is higher compared to the WM-V. Moreover, both tests (WM-F and WM-V) involve the coordination aspect of WM, which is lacking in the DS

backward. Therefore, it is unsurprising that neither measure correlates highly with the DS backward.

Discriminant validity. In terms of discriminant validity, the correlation coefficients between the test procedures and the matrix test correspond to the expected value of approximately $r = .30$ (Conway et al., 2002; Kane et al., 2004; Unsworth, Brewer et al., 2009; Unsworth & Engle, 2005; Wiley et al., 2011).

When comparing verbal analogies and WM tests, the size of the correlation coefficients strongly depends on the test procedure (Unsworth, Brewer et al., 2009). Unsurprisingly, the correlation coefficients for the WM test with a strong verbal orientation (WM-V) were higher than for the purely visual test. As mentioned above, some research suggests that WM can be divided into spatial and non-spatial factors (e.g., Oberauer et al., 2003). The verbal analogies test addresses a verbal facet of WM, while the presented tests mainly capture a visuospatial factor, which would explain the relatively low correlation (see Unsworth, Brewer et al., 2009 for a comparison). This could also be the reason why the two test procedures have differential correlations with verbal analogies despite their great similarity. Furthermore, additional semantic knowledge is necessary for analogical reasoning (Krawczyk et al., 2008; Morrison et al., 2004). Semantic knowledge is also necessary for the WM-V, which might facilitate quick transfer performance when converting the geometric figure into a word. This aspect is not addressed in the WM-F, which might therefore explain the discrepancy in correlations.

Thus, multiple mechanisms could be held responsible for the lower correlation between the verbal analogies test and the proposed WM tests. As Unsworth, Redick et al. (2009) sum up: there is a "'complex' picture of performance in complex WM span tasks

and their relation to measures of higher-order cognition" (Unsworth, Redick et al., 2009, p. 650).

The correlation with the arithmetic test also turned out as expected, at approximately $r = .33$ (e.g., Friso-van den Bos et al., 2013; Peng et al., 2016).

As a comparable composite score to that used by the Bundeswehr has not been examined in the literature, the benchmark was assumed to be around $r = .30$, because the composite score should reflect fluid intelligence, which has a correlation with WM of around $r = .30$ as well (e.g., Unsworth, Brewer et al., 2009). This correlation level was achieved in the present study.

Overall, then, the results with regard to convergent validity were in line with expectations.

Predictive validity. An important factor in the present case is predictive validity. This was evaluated in the present study using ROC analysis. The AUC in psychological diagnostics should range between .70 and .80, while an AUC above .90 indicates design errors rather than excellent diagnostics (Youngstrom, 2014). Particularly among the officer candidates, the AUC in the present case is very good compared to the criteria recommended by Youngstrom (2014). Here again, it can be seen that LLTM and thus all-or-nothing scoring works better than LPCM with partial-credit scoring. The optimal cut-off value was calculated using Youden's index (Youden, 1950), which calculates the optimal score that maximizing specificity and sensitivity. This cut-off seems to be best since the intended purpose of the test is as a "select-in" criterion with a ranking of the participants selected in. Therefore, participants with a score equal to or above the proposed cut-off should be invited for further testing. Since the cut-off is based on the best possible

result in personnel selection, it is no problem selecting in false positives, since it is likely that such persons will pass the personnel selection with excellent results as well. For this reason, the proposed cut-off value should be set at 9 for officer candidates.

In the case of non-commissioned officers, on the other hand, the corresponding values might not sufficiently meet the requirements. This could be the case if the test procedures are very difficult and only a small number of very simple items exist. As a result, there may not be enough variance in the non-commissioned officer candidates to make a sufficiently accurate differentiation. Here, it would probably be appropriate to establish an easier test that would produce greater variance and thus allow a more accurate prediction of performance in on-site diagnostics. If one of the available test procedures were to be chosen nevertheless, it would be better to choose the verbal WM test with all-or-nothing scoring. All-or-nothing scoring is superior in many respects to partial-credit scoring (e.g., in terms of reliability or model fit) and the AUC value of .69 is just below the minimally acceptable value of .70. Therefore, if strictly necessary, the cut-off score for privates, sergeants and non-commissioned officer candidates should be a score of 4 in the WM-V.

As mentioned above, traditional span tasks are not suitable for an above-average ability sample (Draheim et al., 2018); for this reason, a more complex testing approach was chosen. However, this may have backfired since both tests may be too difficult for privates, sergeants and noncommissioned officers and therefore produce insufficient variance to create an appropriate cut-off value for this group of candidates.

Structural validity. In general, applying all-or-nothing scoring vs. partial credit scoring led to small differences in correlations. The largest difference can be seen in the DS backward. This could be because the DS backward is based on partial-credit scoring, meaning that both scores had higher variance, pushing up the correlation coefficient. Therefore, no final conclusion on scoring can be drawn based on the external criteria alone. However, if other considerations are also taken into account, such as the evidence for discriminant and predictive validity, the automatic item generation model fit and the fit to the data, all-or-nothing scoring is clearly preferable.

Reliability. The LLTM also shows better results than the LPCM in terms of reliability. Overall, the reliability can be assessed as medium to high (Fisseni, 1997). The LLTM's higher reliability coefficients show that the LLTM has better fit than the LPCM. Due to a lack of retesting possibilities, it was unfortunately not possible to determine a retest reliability.

Fairness. Due to an insufficient number of female participants, it could not be determined whether the tests show DIF by gender. This should be determined in a further study with a much larger sample.

Limitations

Sample. As described above, the proportion of women in the sample was too small. On the one hand, this prevented some analyses from being carried out; on the other hand, the ratio roughly reflects the Bundeswehr's applicant pool, which also includes few women. It can therefore be assumed that the sample is reasonably representative.

In addition, the same limitations as in Study 1 apply regarding participants' motivation, cognitive preload and assumptions made.

Test criteria. Since it was not possible to validate the tests with respondents who were not applicants for the Bundeswehr, the study design had to be adapted accordingly. For this reason, it was unfortunately not possible to conduct retests and determine retest reliability. Furthermore, the assessment of cognitive ability needed to be short in order to fit into the selection process. For this reason, the tests of discriminant validity had to be already used in the Bundeswehr's personnel selection process thus already planning to be administered. The choice of a test for convergent validity was also limited because it had to be relatively self-explanatory and easy to program.

Test environment. Since the conditions were the same as in Study 1, the limitations concerning the test environment mentioned in Study 1 also apply to Study 2. Furthermore, all officer candidates were tested at around 5 p.m., introducing systematic bias due to the timing of the test.

Another point is that the present study was conducted in a reasonably controlled setting. In later real-world applications, however, the test situation will be much less

controlled, which could in turn have an impact on predictive validity and test quality. There is not much research on discrepancies between OA completed at home and in a personnel selection environment, despite the extensive research on cheating. In a rare example, Xu et al. (2018) examined whether performance on WM tasks was better at home or in the lab. They found slightly better results in the lab. However, internal consistency was better at home. This gives reason to hope that the present study's results with regard to model fit, reliability and validity could be consistent across later applications in less controlled environments.

Test devices. One goal of OA is that the test can run on different end user devices, enabling mobile assessment. For this to be possible, however, it must be ensured that applicants do not accrue disadvantages from using different devices. Display size, for example, could play a role here. While it may not have a huge impact on perceived workload (Hancock, Sawyer, & Stafford, 2015), visual attention is affected by display size (Chen, Liao, & Yeh, 2011). In general, whether a different screen size has an impact on performance very likely depends on the specific test in question (Bridgeman et al., 2003), with WM playing an important role in this context (Arthur, Keiser, Hagen, & Traylor, 2018).

Unfortunately, it was not possible to test and compare different devices like smartphones and tablets against each other in the present study. Before the proposed tests are applied in practice, it should be tested whether equivalent results are achieved on smartphones and tablets compared to computers. In general, computers, laptops and tablets may lead to similar outcomes, while smartphones and phablets may not (Arthur, Keiser, & Doverspike, 2018). In any case, there is a difference between mobile and non-mobile

devices (Morelli, Mahan, & Illingworth, 2014). Thus, until there is evidence that no differences between different devices exist, potential applicants should at least be advised to use a computer for testing.

General Discussion and Conclusion

The aim of the present project was to develop a test for OA in the German Bundeswehr that has a high predictive validity for on-site diagnostics. The specific requirements of OA should be taken into account in the development and design of such a test as well. Therefore, a key issue was whether automatic item generation could be operationally implemented and whether the tests created in this way would have appropriate psychometric properties. In the course of this project, two tests were developed and evaluated. Both were intended to assess WM and were structured similarly to a complex span task. The tests differed primarily in terms of content: while the first test (WM-F) is based solely on figural content, the second test (WM-V) combines verbal and figural content and requires a corresponding transfer performance in this area. For this reason, it can be assumed that it is more difficult to achieve a high score in the WM-V. Since it was not quite clear how long the stimuli should be presented, this hypothesis was also tested in two conditions (1 s vs. 3 s presentation duration).

The 3 s presentation duration seems to be superior to the shorter time, which is why the final test format has a presentation time of 3 s. Both studies demonstrated that automatic item generation worked for the proposed tests. The fit when applying all-or-nothing scoring, and thus applying an LLTM, was so good that it can be considered almost perfect (Baghaei & Hohensinn, 2017). In general, all-or-nothing scoring was found to have superior psychometric properties. However, the main weakness of the present project is that relatively few women participated in both studies. Therefore, it was not possible to analyze, for example, whether there was a significant difference in how men and women processed the test (DIF). It is unlikely that this had an impact on the psychometric properties; nonetheless, it should be mentioned here.

The processing time of both final test forms, at about ten minutes, is quite acceptable for OA and enables fast classification.

As the tests are specifically designed for OA, the risks otherwise associated with OA should at least be minimized (Schaper, 2009). Furthermore, the tests should not be particularly vulnerable to cheating, since automatic item generation makes it impossible to search for answers in the Internet (Steger et al., 2018). The predictive validity for officer candidates in particular is very good, while the other results regarding reliability and validity turned out as expected. The predictive validity for privates, sergeants and non-commissioned officer candidates was less than satisfactory, perhaps because the test is too difficult for this group. It would therefore be advisable that a simpler test be developed that is a better fit for non-commissioned officer candidates.

In summary, it can be said that the WM-F is a test with good to very good psychometric characteristics that is ideally suited for predicting the performance of on-site diagnostics of officer candidates. The use of automatic item generation significantly reduces the risk of cheating resulting from UIT without affecting the psychometric properties of the test.

References

Acikgoz, Y., & Sumer, H. C. (2019). Implementation intentions as a predictor of applicant withdrawal. *Military Psychology*, *31*(5), 347–354. https://doi.org/10.1080/08995605.2019.1637208

Ackerman, P. L., Beier, M. E., & Boyle, M. O. (2002). Individual differences in working memory within a nomological network of cognitive and perceptual speed abilities. *Journal of Experimental Psychology: General*, *131*(4), 567–589. https://doi.org/10.1037//0096-3445.131.4.567

Aguado, D., Vidal, A., Olea, J., Ponsoda, V., Barrada, J. R., & Abad, F. J. (2018). Cheating on unproctored internet test applications: An analysis of a verification test in a real personnel selection context. *The Spanish Journal of Psychology*, *21*, E62. https://doi.org/10.1017/sjp.2018.50

Alexandrowicz, R. W., & Draxler, C. (2016). Testing the Rasch model with the conditional likelihood ratio test: Sample size requirements and bootstrap algorithms. *Journal of Statistical Distributions and Applications*, *3*(2). https://doi.org/10.1186/s40488-016-0039-y

Alloway, T. P., & Alloway, R. G. (2010). Investigating the predictive roles of working memory and IQ in academic attainment. *Journal of Experimental Child Psychology*, *106*, 20–29. https://doi.org/10.1016/j.jecp.2009.11.003

Alloway, T. P., & Alloway, R. G. (2013). Working memory across the lifespan: A cross-sectional approach. *Journal of Cognitive Psychology*, *25*(1), 84–93. https://doi.org/10.1080/20445911.2012.748027

Alvarez, G. A., & Cavanagh, P. (2004). The capacity of visual short-term memory is set both by visual information load and by number of objects. *Psychological Science*, *15*(2), 106–111.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika*, *38*(1), 123–140. https://doi.org/10.1007/BF02291180

Anderson, C. J., Li, Z., & Vermunt, J. K. (2007). Estimation of models in a Rasch family for polytomous items and multiple latent variables. *Journal of Statistical Software*, *20*(6), 1–36.

Anderson, D., Kahn, J. D., & Tindal, G. (2017). Exploring the robustness of a unidimensional Item Response Theory model with empirically multidimensional data. *Applied Measurement in Education*, *30*(3), 163–177. https://doi.org/10.1080/08957347.2017.1316277

Andrich, D. (2010). Sufficiency and conditional estimation of person parameters in the polytomous Rasch model. *Psychometrika*, *75*(2), 292–308. https://doi.org/10.1007/s11336-010-9154-8

Andrich, D. (2016). Georg Rasch and Benjamin Wright's struggle with the unidimensional polytomous model with sufficient statistics. *Educational and Psychological Measurement*, *76*(5), 713–723. https://doi.org/10.1177/0013164416634790

Ansbacher, H. L. (1941). German military psychology. *Psychological Bulletin*, *38*(6), 370–392.

Arendasy, M. E., & Sommer, M. (2010). Evaluating the contribution of different item features to the effect size of the gender difference in three-dimensional mental rotation using automatic item generation. *Intelligence*, *38*(6), 574–581. https://doi.org/10.1016/j.intell.2010.06.004

References

Arendasy, M. E., Sommer, M., & Hergovich, A. (2007). Psychometrische Technologie: Automatische Zwei-Komponenten-Itemgenerierung am Beispiel eines neuen Aufgabentyps zur Messung der Numerischen Flexibilität. *Diagnostica*, *53*(3), 119–130.

Arendasy, M. E., Sommer, M., & Mayr, F. (2011). Using automatic item generation to simultaneously construct German and English versions of a word fluency test. *Journal of Cross-Cultural Psychology*, *43*(3), 464–479. https://doi.org/10.1177/0022022110397360

Arthur, W., Doverspike, D., Muñoz, G. J., Taylor, J. E., & Carr, A. E. (2014). The use of mobile devices in high-stakes remotely delivered assessments and testing. *The International Journal of Selection and Assessment*, *22*(2), 113–123.

Arthur, W., Glaze, R. M., Villado, A. J., & Taylor, J. E. (2009). Unproctored internet-based tests of cognitive ability and personality: Magnitude of cheating and response distortion. *Industrial and Organizational Psychology*, *2*(1), 39–45. https://doi.org/10.1111/j.1754-9434.2008.01105.x

Arthur, W., Keiser, N. L., & Doverspike, D. (2018). An information-processing-based conceptual framework of the effects of unproctored internet-based testing devices on scores on employment-related assessments and tests. *Human Performance*, *31*(1), 1–32.

Arthur, W., Keiser, N. L., Hagen, E., & Traylor, Z. (2018). Unproctored internet-based device-type effects on test scores: The role of working memory. *Intelligence*, *67*, 67–75. https://doi.org/10.1016/j.intell.2018.02.001

Arvey, R. D., Gordon, M. E., & Massengill, D. P. (1975). Differential dropout rates of minority and majority job candidates due to "time lags" between selection procedures. *Personnel Psychology*, *28*, 175–180. https://doi.org/10.1111/j.1744-6570.1975.tb01378.x

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.), *Psychology of Learning and Motivation* (Vol. 2, pp. 89–195). New York, NY: Academic. https://doi.org/10.1016/S0079-7421(08)60422-3

Awh, E., Barton, B., & Vogel, E. K. (2007). Visual working memory represents a fixed number of items regardless of complexity. *Psychological Science*, *18*(7), 622–628.

Awh, E., Jonides, J., Smith, E. E., Schumacher, E. H., Koeppe, R. A., & Katz, S. (1996). Dissociation of storage and rehearsal in verbal working memory: Evidence from Positron Emission Tomography. *Psychological Science*, *7*(1), 25–31.

Babcock, B., & Hodge, K. J. (2020). Rasch versus classical equating in the context of small sample sizes. *Educational and Psychological Measurement*, *80*(3), 499–521. https://doi.org/10.1177/0013164419878483

Baddeley, A. D. (1986a). The central executive and its malfunctions. In A. D. Baddeley (Ed.), *Oxford Psychology Series: Vol. 11. Working memory* (pp. 223–253). Oxford, UK: Oxford University.

Baddeley, A. D. (Ed.) (1986b). *Oxford Psychology Series: Vol. 11*. *Working memory*. Oxford, UK: Oxford University.

Baddeley, A. D. (1990). *Human memory: Theory and practice*. Boston, MA: Allyn & Bacon.

Baddeley, A. D. (1993). Working memory or working attention? In A. D. Baddeley & L. Weiskrantz (Eds.), *Attention, selection, awareness and control: A tribute to Donald Broadbent* (pp. 152–170). Oxford, UK: Oxford University.

Baddeley, A. D. (1996). Exploring the central executive. *Quarterly Journal of Experimental Psychology*, *49A*(1), 5–28.

References

Baddeley, A. D. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, *4*(11).

Baddeley, A. D. (2002). Is working memory still working? *European Psychologist*, *7*(2), 85–97. https://doi.org/10.1027//1016-9040.7.2.85

Baddeley, A. D. (2003). Working memory and language: An overview. *Journal of Communication Disorder*, *36*, 189–208. https://doi.org/10.1016/S0021-9924(03)00019-4

Baddeley, A. D. (2007). *Working memory, thought, and action*. *Oxford Psychology Series: Vol. 45*. New York, NY: Oxford University.

Baddeley, A. D., Chincotta, D., & Adlam, A. (2001). Working memory and the control of action: Evidence from task switching. *Journal of Experimental Psychology: General*, *130*(4), 641–657.

Baddeley, A. D., Grant, S., Wight, E., & Thomson, N. (1975). Imagery and visual working memory. In P. M. A. Rabbitt & S. Dornic (Eds.), *Attention and performance V* (pp. 205–217). London, UK: Academic.

Baddeley, A. D., & Hitch, G. J. (1974). Working memory. In G. H. Bower (Ed.), *The psychology of learning and motivation: Vol. 8. The psychology of learning and motivation: Advances in research and theory* (pp. 47–89). New York: Academic.

Baddeley, A. D., & Lieberman, K. (1980). Spatial working memory. In R. S. Nickerson (Ed.), *Attention and Performance VIII* (pp. 521–539). Hillsdale, NJ: Lawrence Erlbaum Associates.

Baddeley, A. D., & Logie, R. H. (1999). Working memory – The multiple-component model. In A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of*

active maintenance and executive control (pp. 28–61). Cambridge, UK: Cambridge University.

Bae, G.-Y., & Luck, S. J. (2019). What happens to an individual visual working memory representation when it is interrupted? *British Journal of Psychology*, *110*, 268–287. https://doi.org/10.1111/bjop.12339

Baghaei, P., & Hohensinn, C. (2017). A method of q-matrix validation for the Linear Logistic Test Model. *Frontiers in Psychology*, *8*(897). https://doi.org/10.3389/fpsyg.2017.00897

Baghaei, P., & Kubinger, K. D. (2015). Linear Logistic Test Modeling with R. *Practical Assessment, Research & Evaluation*, *20*(1). Retrieved from http://pareonline.net/getvn.asp?v=20&n=1

Baghaei, P., & Ravand, H. (2015). A cognitive processing model of reading comprehension in English as a foreign language using the Linear Logistic Test Model. *Learning and Individual Differences*, *43*, 100–105. https://doi.org/10.1016/j.lindif.2015.09.001

Baghaei, P., Yanagida, T., & Heene, M. (2017). Development of a descriptive fit statistic for the Rasch model. *North American Journal of Psychology*, hprints-01654099.

Baker, F. B. (1993). Sensitivity to the Linear Logistic Test Model to misspecification of the weight matrix. *Applied Psychological Measurement*, *17*(3), 201–210.

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2.th ed.). *Statistics: Vol. 176*. New York, NY: Dekker.

Barbosa, H., & Garcia, F. (2005). Importance of Online assessment in the E-learning process. In ITBHET & IEEE (Ed.), *6th International Conference on Information Technology Based Higher Education and Training* (F3B-1). Santo Domingo, CUB.

References

Bartram, D. (2000). Internet recruitment and selection: Kissing frogs to find princes. *The International Journal of Selection and Assessment*, *8*(4), 261–274. https://doi.org/10.1111/1468-2389.00155

Bayliss, D. M., Jarrold, C., Gunn, D. M., & Baddeley, A. D. (2003). The complexities of complex span: Explaining individual differences in working memory in children and adults. *Journal of Experimental Psychology. General*, *132*(1), 71–92. https://doi.org/10.1037/0096-3445.132.1.71

Bejar, I. I. (1990). A generative analysis of a three-dimensional spatial task. *Applied Psychological Measurement*, *14*(3), 237–245. https://doi.org/10.1177/014662169001400302

Bejar, I. I. (1996). *Generative response modeling: Leveraging the computer as a test delivery medium*. *ETS Research Report: 96-13*. Princeton, NJ: Educational Testing Service.

Bejar, I. I. (2002). Generative testing: From conception to implementation. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 199–217). Mahwah, NJ: Lawrence Erlbaum Associates.

Bejar, I. I., & Cooper, P. F. (2013). *On the feasibility of generating situational judgment tests by means of photorealistic methods* (No. RM-13-08). Princeton, NJ: Educational Testing Service.

Bejar, I. I., Lawless, R. R., Morley, M. E., Wagner, M. E., Bennett, R. E., & Revuelta, J. (2002). *A feasibility study of on-the-fly item generation in adaptive testing* (ETS Research Report No. 02-23). Princeton, NJ. https://doi.org/10.1002/j.2333-8504.2002.tb01890.x

Bensch, D., Maaß, U., Greiff, S., Horstmann, K. T., & Ziegler, M. (2019). The nature of faking: A homogeneous and predictable construct? *Psychological Assessment*, *31*(4), 532–544. https://doi.org/10.1037/pas0000619

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443–459. https://doi.org/10.1007/BF02293801

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement*, *12*(3), 261–280. https://doi.org/10.1177/014662168801200305

Bodmann, S. M., & Robinson, D. H. (2004). Speed and performance differences among computer-based and paper-pencil tests. *Journal of Educational Computing Research*, *31*(1), 51–60.

Bolt, D. M., & Lall, V. F. (2003). Estimation of compensatory and noncompensatory multidimensional Item Response Models using Markov Chain Monte Carlo. *Applied Psychological Measurement*, *27*, 395–414. https://doi.org/10.1177/0146621603258350

Borman, W. C., Klimoski, R. J., & Ilgen, D. R. (2003). Stability and change in industrial and organizational psychology. In W. C. Borman, D. R. Ilgen, & R. J. Klimoski (Eds.), *Handbook of psychology: Industrial and organizational psychology* (pp. 1–17). Hoboken, NJ: John Wiley & Sons.

Bormuth, J. (1969). *On a theory of achievement test items*. Chicago, IL: University of Chicago.

Boss, P., König, C. J., & Melchers, K. G. (2015). Faking good and faking bad among military conscripts. *Human Performance*, *28*, 26–39. https://doi.org/10.1080/08959285.2014.974758

References

Bowman, M. L. (1989). Testing individual differences in ancient China. *American Psychologist*, *44*(3), 576–578. https://doi.org/10.1037/0003-066X.44.3.576.b

Brandt, S. (2012). Robustness of multidimensional analyses against local item dependence. *Psychological Test and Assessment Modeling*, *54*(1), 36.

Breit, M., Brunner, M., & Preckel, F. (2020). General intelligence and specific cognitive abilities in adolescence: Tests of age differentiation, ability differentiation, and their interaction in two large samples. *Developmental Psychology*, *56*(2), 364–384. https://doi.org/10.1037/dev0000876

Bridgeman, B., Lennon, M. L., & Jackenthal, A. (2003). Effects of screen size, screen resolution, and display rate on computer-based test performance. *Applied Measurement in Education*, *16*(3), 191–205. https://doi.org/10.1207/S15324818AME1603_2

Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion* (3rd ed.). *Psychologie*. München, DE: Pearson Studium.

Bühner, M., Kröner, S., & Ziegler, M. (2008). Working memory, visual–spatial-intelligence and their relationship to problem-solving. *Intelligence*, *36*, 672–680. https://doi.org/10.1016/j.intell.2008.03.008

Bundeswehr (2012). *Offizieranwärter bei der Bundeswehr – Die Eignungsprüfung* [video]. Retrieved from https://www.youtube.com/watch?v=ceLc792ZLtA

Bundeswehr (2014). Die Anreise. Retrieved from https://web.archive.org/web/20150524220647/https://mil.bundeswehr-karriere.de/portal/a/milkarriere/!ut/p/c4/DcXBDYAgDAXQWVygvXtzC_VWzA82YCEFJXF6zTs83vln8miUrsUk88rboXMYdGlO4q5wkJ6OgAEPt0WCRvtvHa1TqS-JObSBa1qmD5TQw3I!/

References

Bundeswehr (2016a). Assessment Trainer. Retrieved from http://www.bundeswehrkarriere.de/bewerbung/assessment-trainer

Bundeswehr (2016b). Assessment Trainer: Wortanalogien. Retrieved from http://www.bundeswehrkarriere.de/bewerbung/assessment-trainer/frage6#1502461290550

Bundeswehr (2019a). Assessment Trainer. Retrieved from www.bundeswehrkarriere.de/ihre-berufung/assessment-trainer

Bundeswehr (2019b). Ihr Weg zu uns. Retrieved from https://www.bundeswehrkarriere.de/ihr-weg-zu-uns

Busold, M. (Ed.) (2019). *War for Talents: Erfolgsfaktoren im Kampf um die Besten* (2nd ed.). Berlin, DE: Springer. https://doi.org/10.1007/978-3-662-57481-2

Campbell, J. P. (1990). An overview of the army selection and classification project (Project A). *Personnel Psychology*, *43*, 231–239.

Canty, A. J. (2002). Resampling methods in R: The boot package. *The Newsletter of the R Project*, *2/3*, 2–6.

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review*, *97*(3), 404–431. https://doi.org/10.1037/0033-295X.97.3.404

Carretta, T. R., Perry Jr, D. C., & Ree, M. J. (1996). Prediction of situational awareness in F-15 pilots. *The International Journal of Aviation Psychology*, *6*(1), 21–41.

Carstairs, J., & Myors, B. (2009). Internet testing: A natural experiment reveals test score inflation on a high-stakes, unproctored cognitive test. *Computers in Human Behavior*, *25*(3), 738–742. https://doi.org/10.1016/j.chb.2009.01.011

Case, R., Kurland, D. M., & Goldberg, J. (1982). Operational efficiency and the growth of short-term memory span. *The Journal of Experimental Child Psychology*, *33*(3), 386–404.

Cavanagh, T. (2014). *Cheating on online assessment tests: Prevalence and impact on validity* (Doctoral Dissertation). Colorado State University, Fort Collins, CO.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, *4*, 55–81.

Chen, E. H., & Bailey, D. H. (2020). Dual-task studies of working memory and arithmetic performance: A meta-analysis. *The Journal of Experimental Psychology. Learning, Memory, and Cognition.* Advance online publication. https://doi.org/10.1037/xlm0000822

Chen, I.-P., Liao, C.-N., & Yeh, S.-H. (2011). Effect of display size on visual attention. *Perceptual and Motor Skills*, *112*(3), 959–974. https://doi.org/10.2466/22.24.26.PMS.112.3.959-974

Chen, J., & Perie, M. (2018). Comparability within computer-based assessment: Does screen size matter? *Computers in the Schools*, *35*(4), 268–283. https://doi.org/10.1080/07380569.2018.1531599

Chen, S.-Y., Lei, P.-W., & Liao, W.-H. (2008). Controlling item exposure and test overlap on the fly in computerized adaptive testing. *The British Journal of Mathematical and Statistical Psychology*, *61*, 471–492. https://doi.org/10.1348/000711007X227067

Choi, S. W., Gibbons, L. E., & Crane, P. K. (2016). Package 'lordif'. Retrieved from https://cran.r-project.org/web/packages/lordif/lordif.pdf

Christensen, K. B., Bjorner, J. B., Kreiner, S., & Petersen, J. H. (2002). Testing unidimensionality in polytomous Rasch models. *Psychometrika*, *67*(4), 563–574.

Coifman, K. G., Kane, M. J., Bishop, M., Matt, L. M., Nylocks, K. M., & Aurora, P. (2019). Predicting negative affect variability and spontaneous emotion regulation: Can working memory span tasks estimate emotion regulatory capacity? *Emotion.* Advance online publication. https://doi.org/10.1037/emo0000585

Colom, R., Martínez-Molina, A., Shih, P. C., & Santacreu, J. (2010). Intelligence, working memory, and multitasking performance. *Intelligence*, *38*, 543–551. https://doi.org/10.1016/j.intell.2010.08.002

Colom, R., Rebollo, I., Abad, F. J., & Shih, P. C. (2006). Complex span tasks, simple span tasks, and cognitive abilities: A reanalysis of key studies. *Memory and Cognition*, *34*(1), 158–171. https://doi.org/10.3758/bf03193395

Converse, P. D., Oswald, F. L., Imus, A., Hedricks, C., Roy, R., & Butera, H. (2008). Comparing personality test formats and warnings: Effects on criterion-related validity and test-taker reactions. *The International Journal of Selection and Assessment*, *16*(2), 155–169. https://doi.org/10.1111/j.1468-2389.2008.00420.x

Conway, A. R. A., Cowan, N., Bunting, M. F., Therriault, D. J., & Minkoff, Scott, R. B. (2002). A latent variable analysis of working memory capacity, short-term memory capacity, processing speed, and general fluid intelligence. *Intelligence*, *30*, 163–183. https://doi.org/10.1016/S0160-2896(01)00096-4

Conway, A. R. A., & Engle, R. W. (1996). Individual differences in working memory capacity: More evidence for a general capacity theory. *Memory*, *4*(6), 577–590. https://doi.org/10.1080/741940997

Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review*, *12*(5), 769–786. https://doi.org/10.3758/BF03196772

Cormier, D. C., Bulut, O., McGrew, K. S., & Singh, D. (2017). Exploring the relations between Cattell-Horn-Carroll (CHC) Cognitive Abilities and mathematics achievement. *Applied Cognitive Psychology*, *31*, 530–538. https://doi.org/10.1002/acp.3350

Cowan, N. (1988). Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information-processing system. *Psychological Bulletin*, *104*(2), 163–191.

Cowan, N. (1997). *Attention and memory: An integrated framework*. *Oxford Psychology Series: Vol. 26*. New York, NY: Oxford University.

Cowan, N. (1999). An embedded-processes model of working memory. In A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 62–101). Cambridge, UK: Cambridge University.

Cowan, N. (2000). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*, 87–114.

Cowan, N. (2005). *Working memory capacity*. New York, NY: Taylor & Francis. https://doi.org/10.4324/9780203342398

Cowan, N. (2010). The magical mystery four: How is working memory capacity limited, and why? *Current Directions in Psychological Science*, *19*(1), 51–57. https://doi.org/10.1177/0963721409359277

Cowan, N. (2017). The many faces of working memory and short-term storage. *Psychonomic Bulletin & Review*, *24*(4), 1158–1170. https://doi.org/10.3758/s13423-016-1191-6

Cowan, N., Elliott, E. M., Saults, J. S., Morey, C. C., Mattox, S., Hismjatullina, A., & Conway, A. R. A. (2005). On the capacity of attention: Its estimation and its role in

working memory and cognitive aptitudes. *Cognitive Psychology*, *51*(1), 42–100. https://doi.org/10.1016/j.cogpsych.2004.12.001

Cowan, N., Rouder, J. N., Blume, C. L., & Saults, J. S. (2012). Models of verbal working memory capacity: What does it take to make them work? *Psychological Review*, *119*(3), 480–499. https://doi.org/10.1037/a0027791

Cui, X., Bray, S., Bryant, D. M., Glover, G. H., & Reiss, A. L. (2011). A quantitative comparison of NIRS and fMRI across multiple cognitive tasks. *NeuroImage*, *54*(4), 2808–2821. https://doi.org/10.1016/j.neuroimage.2010.10.069

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, *19*(4), 450–466.

Daneman, M., & Tardif, T. (1987). Working memory and reading skill re-examined. In M. Coltheart (Ed.), *Attention and performance XII: The psychology of reading* (pp. 491–508). London, UK: Routledge.

Davier, M. von (1997). Bootstrapping goodness-of-fit statistics for sparse categorical data: Results of a Monte Carlo study. *Methods of Psychological Research Online*, *2*(2), 29–48.

Davis, G., & Holmes, A. (2005). The capacity of visual short-term memory is not a fixed number of objects. *Memory and Cognition*, *33*(2), 185–195.

Debelak, R. (2018). An evaluation of overall goodness-of-fit tests for the Rasch Model. *Frontiers in Psychology*, *9*, 2710. https://doi.org/10.3389/fpsyg.2018.02710

DeBoeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. *Statistics for social science and public policy*. New York, NY: Springer. Retrieved from http://www.loc.gov/catdir/enhancements/fy0813/2004040850-d.html

Dehn, M. J. (2015). *Essentials of working memory assessment and intervention*. *Essentials of Psychological Assessment Series*. Hoboken, NJ: Wiley.

Diehl, K. A. (1998). *Using cognitive theory and Item Response Theory to extract information from wrong responses* (Unpublished Master's thesis). University of Kansas, Lawrence, KS.

Dirlik, E. M. (2019). The comparison of item parameters estimated from parametric and nonparametric item response theory models in case of the violance of local independence assumption. *International Journal of Progressive Education*, *15*(4), 229–240.

Dix, A., & van der Meer, E. (2015). Arithmetic and algebraic problem solving and resource allocation: The distinct impact of fluid and numerical intelligence. *Psychophysiology*, *52*(4), 544–554. https://doi.org/10.1111/psyp.12367

Dixon, P., LeFevre, J.-A., & Twilley, L. C. (1988). Word knowledge and working memory as predictors of reading skill. *Journal of Educational Psychology*, *80*(4), 465–472. https://doi.org/10.1037/0022-0663.80.4.465

Domínguez, C., López-Cuadrado, J., Armendariz, A., Jaime, A., Heras, J., & Pérez, T. A. (2019). Exploring the differences between low-stakes proctored and unproctored language testing using an Internet-based application. *Computer Assisted Language Learning*, *32*(5-6), 483–509. https://doi.org/10.1080/09588221.2018.1527360

Dorans, N. J., & Kingston, N. M. (1985). The effects of violations of unidimensionality on the estimation of item and ability parameters on item response. *Journal of Educational Measurement*, *22*(4), 249–262. https://doi.org/10.1111/j.1745-3984.1985.tb01062.x

Draheim, C., Harrison, T. L., Embretson, S. E., & Engle, R. W. (2018). What Item Response Theory can tell us about the complex span tasks. *Psychological Assessment*, *30*(1), 116–129. https://doi.org/10.1037/pas0000444

Drasgow, F., Luecht, R. M., & Bennett, R. E. (2006). Technology and testing. In R. L. Brennan (Ed.), *ACE/Praeger series on higher education. Educational measurement* (4th ed., pp. 471–516). Westport, CT: Praeger.

Draxler, C., & Alexandrowicz, R. W. (2015). Sample size determination within the scope of conditional maximum likelihood estimation with special focus on testing the Rasch model. *Psychometrika*, *80*(4), 897–919. https://doi.org/10.1007/s11336-015-9472-y

D'Sa, J. l., Alharbi, M. F., & Visbal-Dionaldo, M. l. (2018). The relationship between item difficulty and non-functioning distractors of multiple choice questions. *International Journal of Nursing Education*, *10*(3), 48. https://doi.org/10.5958/0974-9357.2018.00065.X

DuBois, P. H. (1970). *A history of psychological testing*. Boston, MA: Allyn & Bacon.

Ecker, U. K. H., Oberauer, K., & Lewandowsky, S. (2014). Working memory updating involves item-specific removal. *Journal of Memory and Language*, *74*, 1–15. https://doi.org/10.1016/j.jml.2014.03.006

Edwards, B. D., Franco Watkins, A. M., McAbee, S. T., & Faura, L. (2017). The case for using working memory in practice. *The Industrial-Organizational Psychologist*, *55*(1), 4–7. Retrieved from https://www.siop.org/Research-Publications/TIP/TIP-Back-Issues/2017/July/ArtMID/20297/ArticleID/1576/The-Bridge-Connecting-Science-and-PracticeThe-Case-for-Using-Working-Memory-in-Practice

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association*, *82*(397), 171–185. https://doi.org/10.2307/2289144

Ein-Dor, P. (1971). *Elements of a theory of visual information processing* (Unpublished Doctoral Dissertation). Carnegie-Mellon University, Pittsburgh.

Embretson, S. E. (1984). A general latent trait model for response processes. *Psychometrika*, *49*(2), 175–186.

Embretson, S. E. (1995). A measurement model for linking individual learning to processes and knowledge: Application to mathematical reasoning. *Journal of Educational Measurement*, *32*(3), 277–294. https://doi.org/10.1111/j.1745-3984.1995.tb00467.x

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods*, *3*(3), 380–396.

Embretson, S. E. (1999). Generating items during testing: Psychometric issues and models. *Psychometrika*, *64*(4), 407–433.

Embretson, S. E., & Kingston, N. M. (2018). Automatic item generation: A more efficient process for developing mathematics achievement items? *Journal of Educational Measurement*, *55*(1), 112–131. https://doi.org/10.1111/jedm.12166

Embretson, S. E., & Schneider, L. M. (1989). Cognitive component models for psychometric analogies: Conceptually driven versus interactive process models. *Learning and Individual Differences*, *1*(2), 155–178.

Engle, R. W. (2002). Working memory capacity as executive attention. *Current Directions in Psychological Science*, *11*(1), 19–23. https://doi.org/10.1111/1467-8721.00160

Engle, R. W., & Kane, M. J. (2004). Executive attention, working memory capacity, and a two-factor theory of cognitive control. *The Psychology of Learning and Motivation*, *44*, 145–200.

Engle, R. W., Kane, M. J., & Tuholski, S. W. (1999). Individual differences in working memory capacity and what they tell us about controlled attention, general fluid intelligence, and functions of the prefrontal cortex. In A. Miyake & P. Shah (Eds.),

Models of working memory: Mechanisms of active maintenance and executive control (pp. 102–134). Cambridge, UK: Cambridge University. https://doi.org/10.1017/CBO9781139174909.007

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. A. (1999). Working memory, short-term-memory, and general fluid intelligence: A latent-variable approach. *Journal of Experimental Psychology. General*, *128*(3), 309–331.

Evans, J. J., Floyd, R. G., McGrew, K. S., & Leforgee, M. H. (2002). The relations between measures of Cattell-Horn-Carroll (CHC) Cognitive Abilities and reading achievement during childhood and adolescence. *School Psychology Review*, *31*(2), 246–262.

Farrell, S., Oberauer, K., Greaves, M., Pasiecznik, K., Lewandowsky, S., & Jarrold, C. (2016). A test of interference versus decay in working memory: Varying distraction within lists in a complex span task. *Journal of Memory and Language*, *90*, 66–87. https://doi.org/10.1016/j.jml.2016.03.010

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Field, A. P., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. London, UK: Sage.

Fischer, G. H. (1973). The Linear Logistic Test Model as an instrument in educational research. *Acta Psychologica*, *37*, 359–374.

Fischer, G. H. (1995). Derivations of the Rasch Model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch Models: Foundations, recent developments, and applications* (pp. 15–38). New York, NY: Springer.

Fischer, G. H. (2005). Linear Logistic Test Models. In K. Kempf-Leonard (Ed.), *Encyclopedia of social measurement* (pp. 505–514). Amsterdam, NL: Elsevier. https://doi.org/10.1016/B0-12-369398-5/00453-9

Fischer, G. H., & Ponocny, I. (1994). An extension of the partial credit model with an application to the measurement of change. *Psychometrika*, *59*(2), 177–192.

Fischer, G. H., & Ponocny-Seliger, E. (1998). *Structural Rasch modeling: Handbook of the usage of LPCM-WIN 1.0*. Groningen, NL: ProGAMMA.

Fisseni, H. J. (1997). *Lehrbuch der psychologischen Diagnostik* (2nd ed.). Göttingen, DE: Hogrefe.

Fitts, P. M. (1946). German applied psychology during World War II. *American Psychologist*, *1*, 151–161.

Frey, A., Hartig, J., & Rupp, A. A. (2009). An NCME instructional module on booklet designs in large-scale assessments of student achievement: Theory and practice. *Educational Measurement: Issues and Practice*, *28*(3), 39–53.

Friso-van den Bos, I., van der Ven, S. H. G., Kroesbergen, E. H., & van Luit, J. E. V. (2013). Working memory and mathematics in primary school children: A meta-analysis. *Educational Research Review*, *10*, 29–44.

Fürst, A. J., & Hitch, G. J. (2000). Separate roles for executive and phonological components of working memory in mental arithmetic. *Memory and Cognition*, *28*(5), 774–782.

Galanaki, E. (2002). The decision to recruit online: A descriptive study. *Career Development International*, *7*(4), 243–251.

Ganzach, Y., & Pankaj, P. (2018). Wages, mental abilities and assessments in large scale international surveys: Still not much more than g. *Intelligence*, *69*, 1–7.

Gierl, M. J., & Haladyna, T. M. (Eds.) (2012). *Automatic item generation: Theory and practice*. New York, NY: Routledge.

Gierl, M. J., & Lai, H. (2012a). The role of item models in automatic item generation. *International Journal of Testing*, *12*(3), 273–298. https://doi.org/10.1080/15305058.2011.635830

Gierl, M. J., & Lai, H. (2012b). Using automatic item generation to create items for medical licensure exams. In National Council on Measurement in Education (Chair), *National Council on Measurement in Education,* Vancouver, BC.

Gierl, M. J., & Lai, H. (2016). Automatic item generation. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 410–429). New York, NY: Routledge.

Gierl, M. J., & Lai, H. (2018). Using automatic item generation to create solutions and rationales for computerized formative testing. *Applied Psychological Measurement*, *42*(1), 42–57. https://doi.org/10.1177/0146621617726788

Gierl, M. J., Lai, H., & Turner, S. R. (2012). Using automatic item generation to create multiple-choice test items. *Medical Education*, *46*, 757–765. https://doi.org/10.1111/j.1365-2923.2012.04289.x

Gierl, M. J., Zhou, J., & Alves, C. (2008). Developing a taxonomy of item model types to promote assessment engineering. *The Journal of Technology, Learning and Assessment*, *7*(2). Retrieved from www.jtla.org

Gignac, G. E. (2014). Dynamic mutualism versus g factor theory: An empirical test. *Intelligence*, *42*, 89–97. https://doi.org/10.1016/j.intell.2013.11.004

Gignac, G. E., & Watkins, M. W. (2015). There may be nothing special about the association between working memory capacity and fluid intelligence. *Intelligence*, *52*, 18–23. https://doi.org/10.1016/j.intell.2015.06.006

Giofrè, D., Mammarella, I. C., & Cornoldi, C. (2013). The structure of working memory and how it relates to intelligence in children. *Intelligence*, *41*, 396–406. https://doi.org/10.1016/j.intell.2013.06.006

Glas, C. A. W., & Verhelst, N. D. (1995). Testing the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch Models: Foundations, recent developments, and applications* (pp. 69–95). New York, NY: Springer.

Goeters, K. M., & Lorenz, B. (2002). On the implementation of item generation principles in the design of aptitude testing in aviation. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 339–360). Mahwah, NJ: Lawrence Erlbaum Associates.

Green, K. E., & Smith, R. M. (1987). A comparison of two methods of decomposing item difficulties. *Journal of Educational Statistics*, *12*(4), 369–381.

Greiff, S., Kretzschmar, A., Müller, J. C., Spinath, B., & Martin, R. (2014). The computer-based assessment of complex problem solving and how it is influenced by students' information and communication technology literacy. *Journal of Educational Psychology*, *106*, 666–680. https://doi.org/10.1037/a0035426

Guilleux, A., Blanchin, M., Hardouin, J.-B., & Sébille, V. (2014). Power and sample size determination in the Rasch model: Evaluation of the robustness of a numerical method to non-normality of the latent trait. *PloS One*, *9*(1), e83652. https://doi.org/10.1371/journal.pone.0083652

Gulliksen, H. (1950). *Theory of mental tests*. New York, NY: Wiley.

Guo, J., & Drasgow, F. (2010). Identifying cheating on unproctored internet tests: The z-test and the likelihood ratio test. *The International Journal of Selection and Assessment*, *18*(4), 351–364. https://doi.org/10.1111/j.1468-2389.2010.00518.x

Haberecht, M. F., Menon, V., Warsofsky, I. S., White, C. D., Dyer-Friedman, J., Glover, G. H., . . . Reiss, A. L. (2001). Functional neuroanatomy of visuo-spatial working memory in Turner syndrome. *Human Brain Mapping*, *14*(2), 96–107.

Haberman, S. J. (1977). Maximum likelihood estimates in exponential response models. *The Annals of Statistics*, *5*(5), 815–841. https://doi.org/10.1214/aos/1176343941

Haladyna, T. M. (2012). Automatic item generation: A historical perspective. In M. J. Gierl & T. M. Haladyna (Eds.), *Automatic item generation: Theory and practice* (pp. 13–26). New York, NY: Routledge.

Haladyna, T. M., & Shindoll, R. R. (1989). Item shells: A method for writing effective multiple-choice test items. *Evaluation & the Health Professions*, *12*(1), 97–106. https://doi.org/10.1177/016327878901200106

Hambleton, R. K., & Slater, S. C. (1997). Item Response Theory models and testing practices: Current international status and future directions. *European Journal of Psychological Assessment*, *13*(1), 21–28.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage.

Hambrick, D. Z., & Engle, R. W. (2002). Effects of domain knowledge, working memory capacity, and age on cognitive performance: An investigation of the knowledge-is-power hypothesis. *Cognitive Psychology*, *44*, 339–387.

Hambrick, D. Z., & Engle, R. W. (2003). The role of working memory in problem solving. In J. E. Davidson & R. J. Sternberg (Eds.), *The psychology of problem solving* (pp. 176–206). Cambridge, UK: Cambridge University.

Hancock, D., Sawyer, B. D., & Stafford, S. (2015). The effects of display size on performance. *Ergonomics*, *58*(3), 337–354. https://doi.org/10.1037/t69599-000

Handelsblatt. Militär: Der Bundeswehr fehlt Personal – trotz vieler Bewerber. Retrieved from https://www.handelsblatt.com/politik/deutschland/militaer-der-bundeswehr-fehlt-personal-trotz-vieler-bewerber/23915348.html?ticket=ST-7611344-1NGtaCUUQi5rec0MQgW5-ap3

Harrell, T. W. (1992). Some history of the Army General Classification Test. *Journal of Applied Psychology*, *77*, 875–878.

Harris, K. (2018). Military looks at foreign recruits to boost ranks. Retrieved from https://www.cbc.ca/news/politics/caf-military-foreign-recruits-1.4675889

Hartshorne, J. K. (2008). Visual working memory capacity and proactive interference. *PloS One*, *3*(7), e2716. https://doi.org/10.1371/journal.pone.0002716

Hasl, A., Kretschmann, J., Richter, D., Voelkle, M., & Brunner, M. (2019). Investigating core assumptions of the "American Dream": Historical changes in how adolescents' socioeconomic status, IQ, and GPA are related to key life outcomes in adulthood. *Psychology and Aging*, *34*(8), 1055–1076. https://doi.org/10.1037/pag0000392

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and ltems. *Applied Psychological Measurement*, *9*(2), 139–164.

Hearnshaw, L. S. (1964). *A short history of British psychology: 1840 - 1940*. Westport, CT: Greenwood.

Hedge, C., Oberauer, K., & Leonards, U. (2015). Selection in spatial working memory is independent of perceptual selective attention, but they interact in a shared spatial priority map. *Attention, Perception & Psychophysics*, *77*(8), 2653–2668. https://doi.org/10.3758/s13414-015-0976-4

Heene, M., Draxler, C., Ziegler, M., & Bühner, M. (2011). Performance of the bootstrap Rasch model test under violations of non-intersecting item response functions. *Psychological Test and Assessment Modeling*, *53*(3), 283–294.

Hertel, G., & Konradt, U. (Eds.) (2004). *Internet und Psychologie: Vol. 7. Human resource management im Inter- und Intranet*. Göttingen, DE: Hogrefe.

Hertel, G., Konradt, U., & Orlikowski, B. (2003). Ziele und Strategien von E-Assessment aus Sicht der psychologische Personalauswahl. In U. Konradt & W. Sarges (Eds.), *Schriftenreihe Psychologie für das Personalmanagement: Vol. 21. E-Recruitment und E-Assessment* (pp. 37–53). Göttingen, DE: Hogrefe.

Higgins, D., Futagi, Y., & Deane, P. (2005). *Multilingual generalization of the ModelCreator software for math item generation*. Princeton, NJ.

Hilbert, S., Nakagawa, T. T., Puci, P., Zech, A., & Bühner, M. (2015). The Digit Span Backwards Task: Verbal and visual cognitive strategies in working memory assessment. *European Journal of Psychological Assessment*, *31*, 174–180. https://doi.org/10.1027/1015-5759/a000223

Hoeft, F., Hernandez, A., Parthasarathy, S., Watson, C. L., Hall, S. S., & Reiss, A. L. (2007). Fronto-striatal dysfunction and potential compensatory mechanisms in male adolescents with fragile X syndrome. *Human Brain Mapping*, *28*, 543–554. https://doi.org/10.1002/hbm.20406

Hofmann, W., Gschwendner, T., Friese, M., Wiers, R. W., & Schmitt, M. (2008). Working memory capacity and self-regulatory behavior: Toward an individual differences

perspective on behavior determination by automatic versus controlled processes. *Journal of Personality and Social Psychology*, *95*(4), 962–977. https://doi.org/10.1037/a0012705

Hohensinn, C. (2018). pcIRT: An R Package for polytomous and continuous Rasch Models. *Journal of Statistical Software*, *84*(2), 1–14.

Hornke, L. F., Küppers, A., & Etzel, S. (2000). Konstruktion und Evaluation eines adaptiven Matrizentests. *Diagnostica*, *46*(4), 182–188. https://doi.org/10.1026//0012-1924.46.4.182

Hornke, L. F., & Rettig, K. (1989). Konstruktion eines Tests mit verbalen Analogien (CAT-A2): Weitere Untersuchungen. *Untersuchungen Des Psychologischen Dienstes Der Bundeswehr*, *24*, 49–138.

Hornung, C., Brunner, M., Reuter, R. A. P., & Martin, R. (2011). Children's working memory: Its structure and relationship to fluid intelligence. *Intelligence*, *39*, 210–221.

Huang, H.-Y. (2018). Effects of item calibration errors on computerized adaptive testing under cognitive diagnosis models. *Journal of Classification*, *35*(3), 437–465. https://doi.org/10.1007/s00357-018-9265-y

Illingworth, A. J., Morelli, N. A., Scott, J. C., & Boyd, S. L. (2015). Internet-based, unproctored assessments on mobile and non-mobile devices: Usage, measurement equivalence, and outcomes. *Journal of Business and Psychology*, *30*(2), 325–343. https://doi.org/10.1007/s10869-014-9363-8

International Test Commission (2005). ITC guidelines on computer-based and internet delivered testing. Retrieved from https://www.intestcom.org/files/guideline_computer_based_testing.pdf

Irvine, S. H. (2002). The foundations of item generation for mass testing. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 3–34). Mahwah, NJ: Lawrence Erlbaum Associates.

Irvine, S. H. (2014). *Computerised test generation for cross-national military recruitment*. Amsterdam, NL: IOS. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=806181

Irvine, S. H., Dann, P. L., & Anderson, J. D. (1990). Towards a theory of algorithm-determined cognitive test construction. *British Journal of Psychology*, *81*(2), 173–195. https://doi.org/10.1111/j.2044-8295.1990.tb02354.x

Irvine, S. H., & Kyllonen, P. C. (Eds.) (2002). *Item generation for test development*. Mahwah, NJ: Lawrence Erlbaum Associates.

Jaeggi, S. M., Buschkuehl, M., Perrig, W. J., & Meier, B. (2010). The concurrent validity of the N-back task as a working memory measure. *Memory*, *18*(4), 394–412. https://doi.org/10.1080/09658211003702171

Jäger, A. O., Süß, H.-M., & Beauducel, A. (1997). *Berliner Intelligenzstruktur-Test (BIS Form 4)*. Göttingen, DE: Hogrefe.

Jungholt, T. (2018, January 1). Personalmangel beim Militär: Ursula von der Leyens Nachschub-Illusion. Retrieved from https://www.welt.de/politik/deutschland/plus180602584/Personalmangel-beim-Militaer-Ursula-von-der-Leyens-Nachschub-Illusion.html

Jurecka, A., & Hartig, J. (2007). Computer- und netzwerbasiertes Assessment. In Bundesministerium für Bildung und Forschung (Ed.), *Möglichkeiten und Voraussetzungen technologiebasierter Kompetenzdiagnostik* (pp. 37–48). Bonn, DE.

Kail, R., & Hall, L. K. (2001). Distinguishing short-term memory from working memory. *Memory and Cognition*, *29*(1), 1–9. https://doi.org/10.3758/bf03195735

Kane, M. J., Bleckley, M. K., Conway, A. R. A., & Engle, R. W. (2001). A controlled-attention view of working-memory capacity. *Journal of Experimental Psychology. General*, *130*(2), 169–183.

Kane, M. J., Conway, A. R. A., Hambrick, D. Z., & Engle, R. W. (2007). Variation in working memory capacity as variation in executive attention and control. In A. R. A. Conway, C. Jarrold, M. J. Kane, A. Miyake, & J. N. Towse (Eds.), *Variation in working memory* (pp. 21–48). New York, NY: Oxford University.

Kane, M. J., Conway, A. R. A., Miura, T. K., & Colflesh, G. J. H. (2007). Working memory, attention control, and the N-back task: A question of construct validity. *The Journal of Experimental Psychology. Learning, Memory, and Cognition*, *33*(3), 615–622. https://doi.org/10.1037/0278-7393.33.3.615

Kane, M. J., & Engle, R. W. (2000). Working-memory capacity, proactive interference, and divided attention: Limits on long-term memory retrieval. *The Journal of Experimental Psychology. Learning, Memory, and Cognition*, *26*(2), 336–358.

Kane, M. J., Hambrick, D. Z., & Conway, A. R. A. (2005). Working memory capacity and fluid intelligence are strongly related constructs: Comment on Ackerman, Beier, and Boyle (2005). *Psychological Bulletin*, *131*(1), 66-71;. https://doi.org/10.1037/0033-2909.131.1.66

Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: A latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology. General*, *133*(2), 189–217. https://doi.org/10.1037/0096-3445.133.2.189

Kantrowitz, T. M., & Dainis, A. M. (2014). How secure are unproctored pre-employment tests? Analysis of inconsistent test scores. *Journal of Business and Psychology*, *29*(4), 605–616. https://doi.org/10.1007/s10869-014-9365-6

Kantrowitz, T. M., Dawson, C. R., & Fetzer, M. S. (2011). Computer Adaptive Testing (CAT): A faster, smarter, and more secure approach to pre-employment testing. *Journal of Business and Psychology*, *26*(2), 227–232. https://doi.org/10.1007/s10869-011-9228-3

Katzell, R. A., & Austin, J. T. (1992). From then to now: The development of industrial-organizational psychology in the United States. *Journal of Applied Psychology*, *77*(6), 803–835.

Kemp, C., & Jalbert, A. (2012). *Prototype working memory battery for Canadian Forces personnel selection*. Dubrovnik, HRV. Retrieved from IMTA website: http://www.imta.info/PastConferences/Presentations_v2.aspx?Show=2012

Keppel, G., & Underwood, B. J. (1962). Proactive inhibition in short-term retention of single items. *Journal of Verbal Learning and Verbal Behavior*, *1*, 153–161. https://doi.org/10.1016/S0022-5371(62)80023-1

Kesler, S. R., Haberecht, M. F., Menon, V., Warsofsky, I. S., Dyer-Friedman, J., Neely, E. K., & Reiss, A. L. (2004). Functional neuroanatomy of spatial orientation processing in Turner syndrome. *Cerebral Cortex*, *14*(2), 174–180. https://doi.org/10.1093/cercor/bhg116

Kliegl, R., Smith, J., Heckhausen, J., & Baltes, P. B. (1987). Mnemonic training for the acquisition of skilled digit memory. *Cognition and Instruction*, *4*(4), 203–223. https://doi.org/10.1207/s1532690xci0404_1

Kline, P. (2016). *A handbook of test construction: Introduction to psychometric design*. *Psychology revivals*. London, UK: Routledge.

Koker, M. (2019). British army struggles to recruit soldiers now turns to "millennials". Retrieved from https://www.trtworld.com/opinion/british-army-struggles-to-recruit-soldiers-now-turns-to-millennials-23655

Koller, I., Alexandrowicz, R. W., & Hatzinger, R. (2012). *Das Rasch-Modell in der Praxis: Eine Einführung mit eRm* (Vol. 3786). Wien, AT: Facultas.wuv.

Koller, I., Maier, M. J., & Hatzinger, R. (2015). An empirical power analysis of quasi-exact tests for the Rasch model. *Methodology*, *11*(2), 45–54. https://doi.org/10.1027/1614-2241/a000090

König, C. J., Bühner, M., & Mürling, G. (2005). Working memory, fluid intelligence, and attention are predictors of multitasking performance, but polychronicity and extraversion are not. *Human Performance*, *18*(3), 243–266. https://doi.org/10.1207/s15327043hup1803_3

König, C. J., Klehe, U.-C., Berchtold, M., & Kleinmann, M. (2010). Reasons for being selective when choosing personnel selection procedures. *International Journal of Selection and Assessment*, *18*(1), 17–27. https://doi.org/10.1111/j.1468-2389.2010.00485.x

Konradt, U., Lehmann, K., Böhm-Rupprecht, J., & Hertel, G. (2003). Computer- und internetbasierte Verfahren der Berufseignungsdiagnostik: Ein empirischer Überblick. In U. Konradt & W. Sarges (Eds.), *Schriftenreihe Psychologie für das Personalmanagement: Vol. 21. E-Recruitment und E-Assessment* (pp. 105–124). Göttingen, DE: Hogrefe.

Kosh, A. E., Simpson, M. A., Bickel, L., Kellogg, M., & Sanford-Moore, E. (2019). A cost–benefit analysis of automatic item generation. *Educational Measurement: Issues and Practice*, *38*(1), 48–53. https://doi.org/10.1111/emip.12237

Kosslyn, S. M., Reiser, B. J., Farah, M. J., & Fliegel, S. L. (1983). Generating visual images: Units and relations. *Journal of Experimental Psychology. General*, *112*, 278–303. https://doi.org/10.1037/0096-3445.112.2.278

Krawczyk, D. C., Morrison, R. G., Viskontas, I., Holyoak, K. J., Chow, T. W., Mendez, M. F., . . . Knowlton, B. J. (2008). Distraction during relational reasoning: The role of prefrontal cortex in interference control. *Neuropsychologia*, *46*, 2020–2032. https://doi.org/10.1016/j.neuropsychologia.2008.02.001

Krex, L. (2008). *Studienerfolgsprognose in der Bundeswehr: Evaluation vorhandener und zukünftiger Prädiktoren* (Doctoral Dissertation). Universität Bonn, Bonn, DE.

Krumm, S., Schmidt-Atzert, L., Bühner, M., Ziegler, M., Michalczyk, K., & Arrow, K. (2009). Storage and non-storage components of working memory predicting reasoning: A simultaneous examination of a wide range of ability factors. *Intelligence*, *37*(4), 347–364. https://doi.org/10.1016/j.intell.2009.02.003

Kubinger, K. D. (2009). Applications of the Linear Logistic Test Model in psychometric research. *Educational and Psychological Measurement*, *69*(2), 232–244. https://doi.org/10.1177/0013164408322021

Kubinger, K. D. (2019). Item-Response-Theorie (IRT). Retrieved from https://m.portal.hogrefe.com/dorsch/item-response-theorie-irt/

Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2004). Academic performance, career potential, creativity, and job performance: Can one construct predict them all? *Journal of Personality and Social Psychology*, *86*(1), 148–161. https://doi.org/10.1037/0022-3514.86.1.148

Kupka, K. (2013). Online-Assessments im Recrutainment-Format: Wie gefällt das eigentlich den Bewerbern in der echten Auswahlsituation? In J. Diercks & K. Kupka (Eds.), *Recrutainment: Spielerische Ansätze in Personalmarketing und -auswahl* (Vol.

35, pp. 53–66). Wiesbaden, DE: Springer Gabler. https://doi.org/10.1007/978-3-658-01570-1_4

Kupka, K., Diercks, J., & Kopping, N. (2004). Webbasierte Personalauswahl durch E-Assessment bei Unilever Deutschland. *Wirtschaftspsychologie Aktuell*, *3*, 24–28.

Kurz, R., & Evans, T. (2004). Three generations of on-screen aptitude tests: Equivalence or superiority? In British Psychological Society (Ed.), *British Psychological Society Occupational Psychology Conference Compendium of Abstracts* (p. 202). Leicester, UK.

Kyllonen, P. C. (1996). Is working memory capacity Spearman's g? In I. Dennis & P. Tapsfield (Eds.), *Human abilities: Their nature and measurement* (pp. 49–75). New York, NY: Psychology.

Kyllonen, P. C. (2003). Aptitude testing inspired by information processing: A test of the four-sources model. *The Journal of General Psychology*, *120*(3), 375–405. https://doi.org/10.1080/00221309.1993.9711154

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity?! *Intelligence*, *14*(389-433).

LaDuca, A., Staples, W. I., Templeton, B., & Holzman, G. B. (1986). Item modelling procedure for constructing content-equivalent multiple choice questions. *Medical Education*, *20*(1), 53–56. https://doi.org/10.1111/j.1365-2923.1986.tb01042.x

Lamb, K. (1994). Genetics and Spearman's "g" factor. *Mankind Quarterly*, *34*(4), 379–391.

Lee, K., Lee, M. P., Ang, S. Y., & Stankov, L. (2009). Do measures of working memory predict academic proficiency better than measures of intelligence? *Psychology Science Quarterly*, *51*(4), 403–419.

Lee, K., Ning, F., & Goh, H. C. (2013). Interaction between cognitive and non-cognitive factors: The influences of academic goal orientation and working memory on mathematical performance. *Educational Psychology*, *34*(1), 73–91. https://doi.org/10.1080/01443410.2013.836158

Lee, Y.-W. (2004). Examining passage-related local item dependence (LID) and measurement construct using Q3 statistics in an EFL reading comprehension test. *Language Testing*, *21*(1), 74–100. https://doi.org/10.1191/0265532204lt260oa

Lewandowsky, S., Geiger, S. M., Morrell, D. B., & Oberauer, K. (2010). Turning simple span into complex span: Time for decay or interference from distractors? *The Journal of Experimental Psychology. Learning, Memory, and Cognition*, *36*(4), 958–978. https://doi.org/10.1037/a0019764

Lewis-Peacock, J. A., Kessler, Y., & Oberauer, K. (2018). The removal of information from working memory. *Annals of the New York Academy of Sciences*, *1424*(1), 33–44. https://doi.org/10.1111/nyas.13714

Li, P., Stuart, E. A., & Allison, D. B. (2015). Multiple imputation: A flexible tool for handling missing data. *JAMA*, *314*(18), 1966–1967. https://doi.org/10.1001/jama.2015.15281

Li, X., Xiong, Z., Theeuwes, J., & Wang, B. (2020). Visual memory benefits from prolonged encoding time regardless of stimulus type. *The Journal of Experimental Psychology. Learning, Memory, and Cognition.* Advance online publication. https://doi.org/10.1037/xlm0000847

Lin, X. (2011). *Distributed adaptive e-assessment in a higher education environment* (Doctoral thesis). Buckinghamshire New University, London, UK.

Linacre, J. M. (2004). Rasch model estimation: Further topics. *Journal of Applied Measurement*, *5*(1), 95–110.

Lindqvist, E., & Vestman, R. (2011). The labor market returns to cognitive and noncognitive ability: Evidence from the Swedish enlistment. *American Economic Journal: Applied Economics*, *3*(1), 101–128. https://doi.org/10.1257/app.3.1.101

The Local (2019). Twice as many people to undergo military service tests in Sweden this year. Retrieved from https://www.thelocal.se/20190305/twice-as-many-people-to-be-called-up-to-military-service-in-sweden-this-year

Logie, R. H. (1995). *Visuo-spatial working memory*. East Sussex, UK: Lawrence Erlbaum Associates.

Logie, R. H., Gilhooly, K. J., & Wynn, V. (1994). Counting on working memory in arithmetic problem solving. *Memory and Cognition*, *22*(4), 395–410. https://doi.org/10.3758/BF03200866

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement*, *17*(3), 179–193. https://doi.org/10.1111/j.1745-3984.1980.tb00825.x

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*, 279–281.

Lustig, C., May, C. P., & Hasher, L. (2001). Working memory span and the role of proactive interference. *Journal of Experimental Psychology. General*, *130*(2), 199–207. https://doi.org/10.1037/0096-3445.130.2.199

Lynn, R., Chen, H.-Y., & Chen, Y.-H. (2011). Intelligence in Taiwan: Progressive matrices means and sex differences in means and variances for 6- to 17-year-olds.

Journal of Biosocial Science, *43*(4), 469–474. https://doi.org/10.1017/S0021932010000611

Lynn, R., & Irwing, P. (2004). Sex differences on the progressive matrices: A meta-analysis. *Intelligence*, *32*, 481–498. https://doi.org/10.1016/j.intell.2004.06.008

MacCallum, R. C. (2009). Factor analysis. In A. Maydeu-Olivares & R. E. Millsap (Eds.), *The SAGE handbook of quantitative methods in psychology* (pp. 123–147). London, UK: Sage.

MacDonald, G. T. (2014). *The performance of the Linear Logistic Test Model when the q-matrix is misspecified: A simulation study* (Doctoral dissertation). University of South Florida, Tampa, FL. Retrieved from http://scholarcommons.usf.edu/etd/5065

Magis, D., Yan, D., & Davier, A. A. von (2017). *Computerized adaptive and multistage testing with R: Using packages CatR and MstR*. *Use R! Ser*. Cham, CH: Springer. Retrieved from https://ebookcentral.proquest.com/lib/gbv/detail.action?docID=5161134

Mair, P. (2018). *Modern psychometrics with R*. *Use R!* Cham, CH: Springer. Retrieved from http://www.springer.com

Mair, P., & Hatzinger, R. (2007). CML based estimation of extended Rasch models with the eRm package in R. *Psychology Science*, *49*(1), 26–43.

Mair, P., Hatzinger, R., Maier, M. J., Rusch, T., & Debelak, R. (2019). Package ‘eRm’. Retrieved from ftp://ftp.math.ethz.ch/sfs/pub/Software/R-CRAN/web/packages/eRm/eRm.pdf

Makovski, T. (2016). Does proactive interference play a significant role in visual working memory tasks? *The Journal of Experimental Psychology. Learning, Memory, and Cognition*, *42*(10), 1664–1672. https://doi.org/10.1037/xlm0000262

Makovski, T., & Jiang, Y. V. (2008). Proactive interference from items previously stored in visual working memory. *Memory and Cognition*, *36*(1), 43–52.

Makransky, G., & Glas, C. A. W. (2011). Unproctored internet test verification. *Organizational Research Methods*, *14*(4), 608–630. https://doi.org/10.1177/1094428110370715

Martin-Löf, P. (1973). *Statistiska modeller*, Stockholm, SE.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174. https://doi.org/10.1007/BF02296272

McCants, C. W., Katus, T., & Eimer, M. (2018). The capacity and resolution of spatial working memory and its role in the storage of non-spatial features. *Biological Psychology*, *140*, 108–118. https://doi.org/10.1016/j.biopsycho.2018.12.006

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin*, *114*(3), 449–458.

Meijer, R. R., Sijtsma, K., & Smid, N. G. (1990). Theoretical and empirical comparison of the Mokken and the Rasch approach to IRT measurement. *Applied Psychological Measurement*, *14*(3), 283–298.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*, 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81–97.

Miller, G. A., Galanter, E., & Pribram, K. H. (1960). *Plans and the structure of behavior*. New York, NY: Henry Holt and Co. https://doi.org/10.1037/10039-000

Molenaar, I. W. (1995). Estimation of item parameters. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch Models: Foundations, recent developments, and applications* (pp. 39–51). New York, NY: Springer.

Morelli, N. A., Mahan, R. P., & Illingworth, A. J. (2014). Establishing the measurement equivalence of online selection assessments delivered on mobile versus nonmobile devices. *International Journal of Selection and Assessment*, *22*(2), 124–138. https://doi.org/10.1111/ijsa.12063

Morey, C. C. (2019). Perceptual grouping boosts visual working memory capacity and reduces effort during retention. *British Journal of Psychology*, *110*, 306–327. https://doi.org/10.1111/bjop.12355

Morris, T. P., White, I. R., & Royston, P. (2014). Tuning multiple imputation by predictive mean matching and local residual draws. *BMC Medical Research Methodology*, *14*(75). https://doi.org/10.1186/1471-2288-14-75

Morrison, R. G., Krawczyk, D. C., Holyoak, K. J., Hummel, J. E., Chow, T. W., Miller, B. L., & Knowlton, B. J. (2004). A neurocomputational model of analogical reasoning and its breakdown in frontotemporal lobar degeneration. *Journal of Cognitive Neuroscience*, *16*(2), 260–271. https://doi.org/10.1162/089892904322984553

Murdock, B. B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology*, *64*(5), 482–488.

Nagler, U. K. J., & Witzki, A. (2016). Score-Entwicklung für die Bewertung der Gesamtleistung bei einer Multitaskingaufgabe. In I. Fritsche (Chair), *50. Kongress der Deutschen Gesellschaft für Psychologie,* Leipzig, DE.

Nandakumar, R. (1991). Traditional dimensionality versus essential dimensionality. *Journal of Educational Measurement*, *28*(2), 99–117. https://doi.org/10.1111/j.1745-3984.1991.tb00347.x

Nassar, M. R., Helmers, J. C., & Frank, M. J. (2018). Chunking as a rational strategy for lossy data compression in visual working memory. *Psychological Review*, *125*(4), 486–511. https://doi.org/10.1037/rev0000101

Newman, D. A., Joseph, D. L., & MacCann, C. (2010). Emotional intelligence and job performance: The importance of emotion regulation and emotional labor context. *Industrial and Organizational Psychology*, *3*, 159–164. https://doi.org/10.1111/j.1754-9434.2010.01218.x

O'Neill, T. R., Gregg, J. L., & Peabody, M. R. (2020). Effect of sample size on common item equating using the dichotomous Rasch model. *Applied Measurement in Education*, *33*(1), 10–23. https://doi.org/10.1080/08957347.2019.1674309

Oberauer, K. (2002). Access to information in working memory: Exploring the focus of attention. *The Journal of Experimental Psychology. Learning, Memory, and Cognition*, *28*(3), 411–421. https://doi.org/10.1037//0278-7393.28.3.411

Oberauer, K. (2003). Selective attention to elements in working memory. *Experimental Psychology*, *50*(4), 257–269.

Oberauer, K. (2005a). Binding and inhibition in working memory: Individual and age differences in short-term recognition. *Journal of Experimental Psychology. General*, *134*(3), 368–387. https://doi.org/10.1037/0096-3445.134.3.368

Oberauer, K. (2005b). The measurement of working memory capacity. In O. Wilhelm & R. W. Engle (Eds.), *Handbook of Understanding and Measuring Intelligence* (pp. 393–407). Thousand Oaks, CA: Sage.

Oberauer, K. (2009). Design for a working memory. *Psychology of Learning and Motivation*, *51*, 45–100. https://doi.org/10.1016/S0079-7421(09)51002-X

Oberauer, K. (2018). Removal of irrelevant information from working memory: Sometimes fast, sometimes slow, and sometimes not at all. *Annals of the New York Academy of Sciences*, *1424*(1), 239–255. https://doi.org/10.1111/nyas.13603

Oberauer, K. (2019a). Is rehearsal an effective maintenance strategy for working memory? *Trends in Cognitive Sciences*, *23*(9), 798–809. https://doi.org/10.1016/j.tics.2019.06.002

Oberauer, K. (2019b). Working memory and attention - A conceptual analysis and review. *Journal of Cognition*, *2*(1), 36. https://doi.org/10.5334/joc.58

Oberauer, K. (2019c). Working memory capacity limits memory for bindings. *Journal of Cognition*, *2*(1), 40. https://doi.org/10.5334/joc.86

Oberauer, K., & Eichenberger, S. (2013). Visual working memory declines when more features must be remembered for each object. *Memory and Cognition*, *41*(8), 1212–1227. https://doi.org/10.3758/s13421-013-0333-6

Oberauer, K., Farrell, S., Jarrold, C., & Lewandowsky, S. (2016). What limits working memory capacity? *Psychological Bulletin*, *142*(7), 758–799. https://doi.org/10.1037/bul0000046

Oberauer, K., Lewandowsky, S., Awh, E., Brown, G. D. A., Conway, A. R. A., Cowan, N., . . . Ward, G. (2018). Benchmarks for models of short-term and working memory. *Psychological Bulletin*, *144*(9), 885–958. https://doi.org/10.1037/bul0000153

Oberauer, K., & Lin, H.-Y. (2017). An interference model of visual working memory. *Psychological Review*, *124*(1), 21–59. https://doi.org/10.1037/rev0000044

Oberauer, K., Schulze, R., Wilhelm, O., & Süß, H.-M. (2005). Working memory and intelligence—Their correlation and their relation: Comment on Ackerman, Beier, and Boyle (2005). *Psychological Bulletin*, *131*(1), 61–65.

Oberauer, K., Süß, H.-M., Schulze, R., Wilhelm, O., & Wittmann, W. W. (2000). Working memory capacity – Facets of a cognitive ability construct. *Personality and Individual Differences*, *29*(6), 1017–1045.

Oberauer, K., Süß, H.-M., Wilhelm, O., & Wittmann, W. W. (2003). The multiple faces of working memory: Storage, processing, supervision, and coordination. *Intelligence*, *31*, 167–193. https://doi.org/10.1016/S0160-2896(02)00115-0

Oberauer, K., Süß, H.-M., Wilhelm, O., & Wittmann, W. W. (2008). Which working memory functions predict intelligence? *Intelligence*, *36*, 641–652. https://doi.org/10.1016/j.intell.2008.01.007

O'Donnell, R. E., Clement, A., & Brockmole, J. R. (2018). Semantic and functional relationships among objects increase the capacity of visual working memory. *The Journal of Experimental Psychology. Learning, Memory, and Cognition*, *44*(7), 1151–1158. https://doi.org/10.1037/xlm0000508

Oettershagen, K. (2015). *The CAT5 System.* International Military Testing Association, Stockholm, SE.

Olson, I. R., & Jiang, Y. V. (2002). Is visual short-term memory object based? Rejection of the "strong-object" hypothesis. *Perception & Psychophysics*, *64*(7), 1055–1067.

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Pelt, D. H.M., van der Linden, D., & Born, M. P. (2018). How emotional intelligence might get you the job: The relationship between trait emotional intelligence and faking on personality tests. *Human Performance*, *31*(1), 33–54. https://doi.org/10.1080/08959285.2017.1407320

Peng, P., Namkung, J., Barnes, M., & Sun, C. (2016). A meta-analysis of mathematics and working memory: Moderating effects of working memory domain, type of mathematics skill, and sample characteristics. *Journal of Educational Psychology*, *108*(4), 455–473. https://doi.org/10.1037/edu0000079

Piotrowski, C., & Armstrong, T. (2006). Current recruitment and selection practices: A national survey of Fortune 1000 firms. *North American Journal of Psychology*, *8*(3), 489–496.

Ponocny, I. (2001). Nonparametric goodness-of-fit tests for the Rasch model. *Psychometrika*, *66*(3), 437–459.

Postle, B. R., & Brush, L. N. (2004). The neural bases of the effects of item-nonspecific proactive interference in working memory. *Cognitive, Affective, & Behavioral Neuroscience*, *4*(3), 379–392.

Postle, B. R., Brush, L. N., & Nick, A. M. (2004). Prefrontal cortex and the mediation of proactive interference in working memory. *Cognitive, Affective, & Behavioral Neuroscience*, *4*(4), 600–608.

Potosky, D., & Bobko, P. (2004). Selection testing via the internet: Practical considerations and exploratory empirical findings. *Personnel Psychology*, *57*, 1003–1034. https://doi.org/10.1111/j.1744-6570.2004.00013.x

Qian, J., Zhang, K., Liu, S., & Lei, Q. (2019). The transition from feature to object: Storage unit in visual working memory depends on task difficulty. *Memory and Cognition*, *47*(8), 1498–1514. https://doi.org/10.3758/s13421-019-00956-y

Rasch, G. (1960). *Probabilistic model for some intelligence and achievement tests*. Copenhagen, DK: Danish Institute for Educational Research.

Raven, J. (1981). *Manual for Raven's Progressive Matrices and Vocabulary Scales*. Oxford, UK: Oxford University.

Raven, J., Raven, J. C., & Court, J. H. (2003). *Manual for Raven's Progressive Matrices and Vocabulary Scales*. Oxford, UK: Oxford University.

Raven, J. C., Court, J. H., & Raven, J. (2008). *Raven's Standard Progressive Matrices and Vocabulary Scales*. London, UK: Pearson Assessment.

Reckase, M. D. (2009). *Multidimensional item response theory*. Berlin, DE: Springer.

Redick, T. S., & Lindsey, D. R. B. (2013). Complex span and n-back measures of working memory: A meta-analysis. *Psychonomic Bulletin & Review*, *20*, 1102–1113. https://doi.org/10.3758/s13423-013-0453-9

Ree, M. J., & Earles, J. A. (1992). Intelligence is the best predictor of job performance. *Current Directions in Psychological Science*, *1*(3), 86–89.

Ree, M. J., Earles, J. A., & Teachout, M. S. (1994). Predicting job performance: Not much more than g. *Journal of Applied Psychology*, *79*(4), 518–524.

Reeves, T. C. (2000). Alternative assessment approaches for online learning environments in higher education. *Journal of Educational Computing Research*, *23*(1), 101–111.

Rentz, R., & Bashaw, W. (1977). The National Reference Scale for Reading: An application of the Rasch model. *Journal of Educational Measurement*, *14*(2), 161–179.

Repovš, G., & Baddeley, A. D. (2006). The multi-component model of working memory: Explorations in experimental cognitive psychology. *Neuroscience*, *139*, 5–21.

Rey-Mermet, A., Gade, M., Souza, A. S., Bastian, C. C. von, & Oberauer, K. (2019). Is executive control related to working memory capacity and fluid intelligence? *Journal of Experimental Psychology. General*, *148*(8), 1335–1372. https://doi.org/10.1037/xge0000593

Ridgeway, J., McCusker, S., & Pead, D. (2004). *Literature review of e-assessment* (Futurelab No. 10). Bristol, UK.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). Proc: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*, 12–77. https://doi.org/10.1186/1471-2105-12-77

Rosen, V. M., & Engle, R. W. (1998). Working memory capacity and suppression. *Journal of Memory and Language*, *39*, 418–436.

Rost, J. (2001). The growing family of Rasch models. In A. Boomsma, M. A. J. Duijn, & T. A. B. Snijders (Eds.), *Lecture Notes in Statistics: Vol. 157. Essays on Item Response Theory* (Vol. 157, pp. 25–42). New York, NY: Springer. https://doi.org/10.1007/978-1-4613-0169-1_2

Rost, J. (2004). *Lehrbuch Testtheorie - Testkonstruktion* (2.th ed.). *Psychologie Lehrbuch*. Bern, CHE: Huber.

Rost, J., & Spada, H. (1982). Probabilistische Testtheorie. In K. J. Klauer (Ed.), *Schwann Studienbücher. Handbuch der pädagogischen Diagnostik* (pp. 59–98). Düsseldorf, DE: Schwann.

Roulin, N., & Powell, D. M. (2018). Identifying applicant faking in job interviews. *Journal of Personnel Psychology*, *17*(3), 143–154. https://doi.org/10.1027/1866-5888/a000207

Rudner, L. (2010). Implementing the graduate management admission test computerised adaptive test. In W. J. van der Linden & C. A. W. Glas (Eds.), *Statistics for Socal and Behavioral Sciences. Elements of adaptive testing* (pp. 151–165). New York, NY: Springer.

Salgado, J. F. (2001). Some landmarks of 100 years of scientific personnel selection at the beginning of the new century. *The International Journal of Selection and Assessment*, *9*, 3–8.

Salgado, J. F., Anderson, N. R., & Hülsheger, U. R. (2010). Employee selection in Europe: Psychotechnics and the forgotten history of modern scientific employee selection. In J. L. Fart & N. T. Tippins (Eds.), *Handbook of employee selection* (pp. 921–941). New York, NY: Routledge.

Salthouse, T. A., & Mitchell, D. R. D. (1989). Structural and operational capacities in integrative spatial ability, *4*(1), 18–25.

Saylik, R., Raman, E., & Szameitat, A. J. (2018). Sex differences in emotion recognition and working memory tasks. *Frontiers in Psychology*, *9*, 1072. https://doi.org/10.3389/fpsyg.2018.01072

Schaper, N. (2009). Online-Tests aus diagnostisch-methodischer Sicht. In H. Steiner (Ed.), *Online-Assessment: Grundlagen und Anwendung von Online-Tests in der Unternehmenspraxis* (1st ed., pp. 17–36). Heidelberg, DE: Springer. https://doi.org/10.1007/978-3-540-78919-2_2

Scherbaum, C. A., Goldstein, H. W., Yusko, K. P., Ryan, R., & Hanges, P. J. (2012). Intelligence 2.0: Reestabilishing a research program on g in I-O Psychology. *Industrial and Organizational Psychology*, *5*, 128–148.

Schmeichel, B. J., & Demaree, H. A. (2010). Working memory capacity and spontaneous emotion regulation: High capacity predicts self-enhancement in response to negative feedback. *Emotion*, *10*(5), 739–744. https://doi.org/10.1037/a0019355

Schmeiser, C. B., & Welch, C. J. (2006). Test development. In R. L. Brennan (Ed.), *ACE/Praeger series on higher education. Educational measurement* (4th ed., pp. 307–353). Westport, CT: Praeger.

Schmidt, F. L., & Hunter, J. E. (1981). Employment testing: Old theories and new research findings. *American Psychologist*, *36*(10), 1128–1137. https://doi.org/10.1037//0003-066X.36.10.1128

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, *124*(2), 262–274.

Schmidt, F. L., & Hunter, J. E. (2000). Select on intelligence. In E. A. Locke (Ed.), *Handbooks in management. The Blackwell handbook of principles of organizational behavior* (pp. 1–14). Hoboken, NJ: Blackwell.

Schmidt, F. L., Hunter, J. E., & Outerbridge, A. N. (1986). Impact of job experience and ability on job knowledge, work sample performance, and supervisory ratings of job performance. *Journal of Applied Psychology*, *71*(3), 432–439. https://doi.org/10.1037/0021-9010.71.3.432

Schmidt, F. L., Oh, I.-S., & Shaffer, J. A. (2016). *The validity and utility of selection methods in personnel psychology: Practican and theoretical implications of 100 years…* (No. Working Paper). Retrieved from http://home.ubalt.edu/tmitch/645/session%204/Schmidt%20&%20Oh%20MKUP%20validity%20and%20util%20100%20yrs%20of%20research%20Wk%20PPR%202016.pdf

Schmiedek, F., Hildebrandt, A., Lövdén, M., Wilhelm, O., & Lindenberger, U. (2009). Complex span versus updating tasks of working memory: The gap is not that deep. *The Journal of Experimental Psychology. Learning, Memory, and Cognition*, *35*(4), 1089–1096. https://doi.org/10.1037/a0015730

Schneider, W. J., & McGrew, K. S. (2012). The Cattell-Horn-Carroll Model of Intelligence. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed.). New York, NY: Guilford.

Schneider, W. J., & McGrew, K. S. (2018). The Cattell-Horn-Carroll Theory of Cognitive Abilities. In D. P. Flanagan, E. M. McDonough, & A. S. Kaufman (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (4th ed., pp. 73–163). New York, NY: Guilford.

Schroeders, U., & Wilhelm, O. (2010). Testing reasoning ability with handheld computers, notebooks, and paper and pencil. *European Journal of Psychological Assessment*, *26*(4), 284–292. https://doi.org/10.1027/1015-5759/a000038

Schroeders, U., Wilhelm, O., & Schipolowski, S. (2010). Internet-based ability testing: Problems and opportunities. In S. D. Gosling & J. A. Johnson (Eds.), *Advanced methods for conducting online behavioral research* (1st ed., pp. 131–148). Washington, D.C.: American Psychological Association.

Shah, P., & Miyake, A. (1996). The separability of working memory resources for spatial thinking and language processing: An individual differences approach. *Journal of Experimental Psychology. General*, *125*(1), 4–27. https://doi.org/10.1037/0096-3445.125.1.4

Shepherdson, P., Oberauer, K., & Souza, A. S. (2018). Working memory load and the retro-cue effect: A diffusion model account. *The Journal of Experimental Psychology. Human Perception and Performance*, *44*(2), 286–310. https://doi.org/10.1037/xhp0000448

Shoval, R., Luria, R., & Makovski, T. (2019). Bridging the gap between visual temporary memory and working memory: The role of stimuli distinctiveness. *The Journal of*

Experimental Psychology. Learning, Memory, and Cognition. Advance online publication. https://doi.org/10.1037/xlm0000778

Singley, M. K., & Bennett, R. E. (2002). Item generation and beyond: Applications of schema theory to mathematics assessment. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 361–384). Mahwah, NJ: Lawrence Erlbaum Associates.

Sinharay, S., & Johnson, M. S. (2012). Statistical modeling of automatically generated items. In M. J. Gierl & T. M. Haladyna (Eds.), *Automatic item generation: Theory and practice* (183-195). New York, NY: Routledge.

Skaggs, G., & Lissitz, R. W. (1986). An exploration of the robustness of four test equating models. *Applied Psychological Measurement*, *10*(3), 303–317. https://doi.org/10.1177/014662168601000308

Slinde, J. A., & Linn, R. L. (1978). An exploration of the adequacy of the Rasch model for the problem of vertical equating. *Journal of Educational Measurement*, *15*(1), 23–35.

Smith, E. E., & Jonides, J. (1997). Working memory: A view from neuroimaging. *Cognitive Psychology*, *33*(1), 5–42. https://doi.org/10.1006/cogp.1997.0658

Smith, E. V. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement*, *3*(2), 205–231.

Son, G., Oh, B.-I., Kang, M.-S., & Chong, S. C. (2020). Similarity-based clusters are representational units of visual working memory. *The Journal of Experimental Psychology. Learning, Memory, and Cognition*, *46*(1), 46–59. https://doi.org/10.1037/xlm0000722

Sonnleitner, P. (2008). Using the LLTM to evaluate an item-generating system for reading comprehension. *Psychology Science Quarterly*, *50*(3), 345–362.

Spearman, C. (1923). *The nature of" intelligence" and the principles of cognition*. London, UK: Macmillan.

Spearman, C. (1927). *The abilities of man: Their nature and measurement*. New York, NY: Macmillan.

Sprung, L., & Sprung, H. (2001). History of modern psychology in Germany in 19th and 20th century thought and society. *International Journal of Psychology*, *36*, 364–376.

Squires, N. (2019). Italian army struggles to find enough recruits as cosseted millennials find military life too tough. Retrieved from https://www.telegraph.co.uk/news/2019/05/16/italian-army-struggles-find-enough-recruits-cosseted-millennials/

Statistisches Bundesamt (2019). Computer- und Internetnutzung im ersten Quartal des jeweiligen Jahres von Personen ab 10 Jahren. Retrieved from https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Einkommen-Konsum-Lebensbedingungen/IT-Nutzung/Tabellen/zeitvergleich-computernutzung-ikt.html

Steger, D., Schroeders, U., & Gnambs, T. (2018). A meta-analysis of test scores in proctored and unproctored ability assessments. *European Journal of Psychological Assessment*, *27*, 1–11. https://doi.org/10.1027/1015-5759/a000494

Steiner, H. (2017). Online-Assessments als zukünftiger fest integrierter Bestandteil für die Führungskräfteauswahl. In C. von Au (Ed.), *Leadership und Angewandte Psychologie. Auswahl und Onboarding von Führungspersönlichkeiten: Diagnose, Assessment und Integration* (pp. 131–144). Wiesbaden, DE: Springer Fachmedien.

Sternberg, R. J. (1977). Component processes in analogical reasoning. *Psychological Review*, *84*(4), 353–378. https://doi.org/10.1037/0033-295X.84.4.353

Steyer, R., & Partchev, I. (2000). Latent state-trait theory in computerized adaptive testing. In International Military Testing Association (Ed.), *Proceedings of the 42nd Annual Conference of The International Military Testing Association* (pp. 50–55). Edinburgh, UK.

Stigler, S. M. (1999). *Statistics on the table: The history of statisical concepts and methods*. Cambridge, MA: Harvard University.

Süß, H.-M., Oberauer, K., Wittmann, W. W., Wilhelm, O., & Schulze, R. (2002). Working-memory capacity explains reasoning ability—and a little bit more. *Intelligence*, *30*(3), 261–288. https://doi.org/10.1016/S0160-2896(01)00100-3

Szmalec, A., Verbruggen, F., Vandierendonck, A., & Kemps, E. (2011). Control of interference during working memory updating. *Journal of Experimental Psychology: Human Perception and Performance*, *37*, 137–151.

Terman, L. M. (1916). *The measurement of intelligence: An explanation of and a complete guide for the use of the Stanford revision and extension of the Binet-Simon intelligence scale*. Boston, MA: Houghton Mifflin.

Thalmann, M., & Oberauer, K. (2017). Domain-specific interference between storage and processing in complex span is driven by cognitive and motor operations. *Quarterly Journal of Experimental Psychology*, *70*(1), 109–126. https://doi.org/10.1080/17470218.2015.1125935

Thalmann, M., Souza, A. S., & Oberauer, K. (2019). How does chunking help working memory? *The Journal of Experimental Psychology. Learning, Memory, and Cognition*, *45*(1), 37–55. https://doi.org/10.1037/xlm0000578

Thigpen, N., Petro, N. M., Oschwald, J., Oberauer, K., & Keil, A. (2019). Selection of visual objects in perception and working memory one at a time. *Psychological Science*, *30*(9), 1259–1272. https://doi.org/10.1177/0956797619854067

Tippins, N. T., Beaty, J., Drasgow, F., Gibson, W. M., Pearlman, K., Segall, D. O., & Shepherd, W. (2006). Unproctored internet testing in employment settings. *Personnel Psychology*, *59*(1), 189–225.

Turner, M. L., & Engle, R. W. (1989). Is working memory capacity task dependent? *Journal of Memory and Language*, *28*(2), 127–154. https://doi.org/10.1016/0749-596X(89)90040-5

Unsworth, N. (2010). Interference control, working memory capacity, and cognitive abilities: A latent variable analysis. *Intelligence*, *38*(2), 255–267. https://doi.org/10.1016/j.intell.2009.12.003

Unsworth, N., Brewer, G. A., & Spillers, G. J. (2009). There's more to the working memory capacity-fluid intelligence relationship than just secondary memory. *Psychonomic Bulletin & Review*, *16*(5), 931–937.

Unsworth, N., & Engle, R. W. (2005). Working memory capacity and fluid abilites: Examining the correlation between Operation Span and Raven. *Intelligence*, *33*, 67–81.

Unsworth, N., & Engle, R. W. (2007). On the division of short-term and working memory: An examination of simple and complex span and their relation to higher order abilities. *Psychological Bulletin*, *133*(6), 1038–1066. https://doi.org/10.1037/0033-2909.133.6.1038

Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods*, *37*(3), 498–505. https://doi.org/10.3758/BF03192720

Unsworth, N., Redick, T. S., Heitz, R. P., Broadway, J. M., & Engle, R. W. (2009). Complex working memory span tasks and higher-order cognition: A latent-variable analysis of the relationship between processing and storage. *Memory*, *17*(6), 635–654. https://doi.org/10.1080/09658210902998047

Van der Linden, W. J., & van Krimpen-Stoop, E. M. L. A. (2003). Using response times to detect aberrant responses in computerized adaptive testing. *Psychometrika*, *68*(2), 251–265. https://doi.org/10.1007/BF02294800

Vandierendonck, A. (2012). Role of working memory in task switching. *Psychologica Belgica*, *52*(2-3), 229–253.

Vandierendonck, A., Liefooghe, B., & Verbruggen, F. (2010). Task switching: Interplay of reconfiguration and interference control. *Psychological Bulletin*, *136*(4), 601–626. https://doi.org/10.1037/a0019791

Verguts, T., & Boeck, P. de (2000). A note on the Martin-Löf test for unidimensionality. *Methods of Psychological Research Online*, *5*(1), 77–82.

Vinchur, A. J., & Koppes Bryan, L. L. (2012). A history of personnel selection and assessment. In N. Schmitt (Ed.), *The Oxford Handbook of Personnel Assessment and Selection* (pp. 9–30). Oxford, UK: Oxford University.

Voyer, D., Voyer, S. D., & Saint-Aubin, J. (2017). Sex differences in visual-spatial working memory: A meta-analysis. *Psychonomic Bulletin & Review*, *24*, 307–334. https://doi.org/10.3758/s13423-016-1085-7

Wagner, A., & Klein, F. (2015). *Usability evaluation of Computer-Assisted-Testing Software (CAT5).* International Military Testing Association, Stockholm, SE. Retrieved from www.imta.info/PastConferences/Presentations.aspx

Wagner, S. (2006). *Kombinatorik*, University Graz, Graz, AT.

Wainer, H. (2002). On the automatic generation of test items: Some whens, whys, and hows. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 287–305). Mahwah, NJ: Lawrence Erlbaum Associates.

Wang, W.-C., & Wilson, M. (2005). Exploring local item dependence using a random-effects facet model. *Applied Psychological Measurement*, *29*(4), 296–318. https://doi.org/10.1177/0146621605276281

Waschl, N. A., Nettelbeck, T., & Burns, N. R. (2017). The role of visuospatial ability in the Raven's Progressive Matrices. *Journal of Individual Differences*, *38*(4), 241–255. https://doi.org/10.1027/1614-0001/a000241

Wechsler, D. (1993). *Wechsler objective reading dimensions*. London, UK: Pearson Assessment.

Wechsler, D. (1996). *Wechsler objective numerical dimensions*. London, UK: Pearson Assessment.

Wechsler, D. (2008). *Wechsler adult intelligence scale–Fourth Edition (WAIS–IV)*. San Antonio, TX: NCS Pearson.

Whitely, S. E. (1976). Solving verbal analogies: Some cognitive components of intelligence test items. *Journal of Educational Psychology*, *68*(2), 234–242. https://doi.org/10.1037/0022-0663.68.2.234

Whitely, S. E., & Schneider, L. M. (1981). Information structure for geometric analogies: A test theory approach. *Applied Psychological Measurement*, *5*(3), 383–397.

Wickens, D. D. (1973). Some characteristics of word encoding. *Memory and Cognition*, *1*(4), 485–490.

Wickens, D. D., Born, D. G., & Allen, C. K. (1963). Proactive inhibition and item similarity in short-term memory. *Journal of Verbal Learning and Verbal Behavior*, *2*(5-6), 440–445. https://doi.org/10.1016/S0022-5371(63)80045-6

Wiedmann, J. (2009). Mehrstufiges Auswahlverfahren mit Online-Assessments bei der Lufthansa. In H. Steiner (Ed.), *Online-Assessment: Grundlagen und Anwendung von Online-Tests in der Unternehmenspraxis* (1st ed., pp. 105–126). Heidelberg, DE: Springer.

Wiley, J., & Jarosz, A. F. (2012). Working memory capacity, attentional focus, and problem solving. *Current Directions in Psychological Science*, *21*(4), 258–262. https://doi.org/10.1177/0963721412447622

Wiley, J., Jarosz, A. F., Cushen, P. J., & Colflesh, G. J. H. (2011). New rule use drives the relation between working memory capacity and Raven's Advanced Progressive Matrices. *The Journal of Experimental Psychology. Learning, Memory, and Cognition*, *37*(1), 256–263.

Wilhelm, O., & McKnight, P. E. (2002). Ability and achievement testing on the world wide web. In B. Batinic, U.-D. Reips, & M. Bosnjak (Eds.), *Online social sciences* (pp. 167–193). Seattle, WA: Hogrefe & Huber.

Wittmann, W. W., & Süß, H.-M. (1999). Investigating the paths between working memory, intelligence, knowledge, and complex problem-solving performances via Brunswik symmetry. In P. L. Ackerman, P. C. Kyllonen, & R. D. Roberts (Eds.), *Learning and individual differences: Process, trait, and content determinants* (pp. 77–108). Washington, D.C.: American Psychological Association. https://doi.org/10.1037/10315-004

Wolfgang, B. Military struggles to recruit best, brightest in booming economy. Retrieved from https://www.washingtontimes.com/news/2019/jul/7/army-misses-recruiting-goal-booming-economy/

Wolgers, G. (2015). *Military pilot selection process in the Swedish Armed Forces.* International Military Testing Association, Stockholm, SE. Retrieved from www.imta.info/PastConferences/Presentations.aspx

Wright, B., & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement*, *29*, 23–48. https://doi.org/10.1177/001316446902900102

Wrulich, M., Brunner, M., Stadler, G., Schalke, D., Keller, U., & Martin, R. (2014). Forty years on: Childhood intelligence predicts health in middle adulthood. *Health Psychology : Official Journal of the Division of Health Psychology, American Psychological Association*, *33*(3), 292–296. https://doi.org/10.1037/a0030727

Xu, G., Wang, C., & Shang, Z. (2016). On initial item selection in cognitive diagnostic computerized adaptive testing. *The British Journal of Mathematical and Statistical Psychology*, *69*(3), 291–315. https://doi.org/10.1111/bmsp.12072

Xu, Z., Adam, K. C. S., Fang, X., & Vogel, E. K. (2018). The reliability and stability of visual working memory capacity. *Behavior Research Methods*, *50*(2), 576–588. https://doi.org/10.3758/s13428-017-0886-6

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*(3), 187–213. https://doi.org/10.1111/j.1745-3984.1993.tb00423.x

Yerkes, R. M. (Ed.) (1921). *Memoirs of the National Academy of Sciences: XV*. *Psychological examining in the United States Army*. Washington, D.C.: Washington Government Printing Office.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, *3*, 32–35. https://doi.org/10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3

Youngstrom, E. A. (2014). A primer on receiver operating characteristic analysis and diagnostic efficiency statistics for pediatric psychology: We are ready to ROC. *Journal of Pediatric Psychology*, *39*(2), 204–221. https://doi.org/10.1093/jpepsy/jst062

Zhang, Q., Shen, M., Tang, N., Zhao, G., & Gao, Z. (2013). Object-based encoding in visual working memory: A life span study. *Journal of Vision*, *13*(10). https://doi.org/10.1167/13.10.11

Ziegler, M. (2014). Stop and state your intentions! *European Journal of Psychological Assessment*, *30*(4), 239–242. https://doi.org/10.1027/1015-5759/a000228

Ziegler, M., & Bensch, D. (2013). Lost in translation: Thoughts regarding the translation of existing psychological measures into other languages. *European Journal of Psychological Assessment*, *29*(2), 81–83. https://doi.org/10.1027/1015-5759/a000167

Ziegler, M., & Brunner, M. (2016). Test standards and psychometric modeling. In A. A. Lipnevich, F. Preckel, & R. D. Roberts (Eds.), *The Springer Series on Human Exceptionality. Psychosocial skills and school systems in the 21st century: Theory, research, and practice* (pp. 29–55). Cham, CH: Springer International Publishing. https://doi.org/10.1007/978-3-319-28606-8_2

Ziegler, M., Dietl, E., Danay, E., Vogel, M., & Bühner, M. (2011). Predicting training success with general mental ability, specific ability tests, and (un)structured interviews: A meta-analysis with unique samples. *The International Journal of Selection and Assessment*, *19*(2), 170–182. https://doi.org/10.1111/j.1468-2389.2011.00544.x

Ziegler, M., & Hagemann, D. (2015). Testing the unidimensionality of items. *European Journal of Psychological Assessment*, *31*(4), 231–237. https://doi.org/10.1027/1015-5759/a000309

Ziegler, M., Maaß, U., Griffith, R., & Gammon, A. (2015). What is the nature of faking? Modeling distinct response patterns and quantitative differences in faking at the same time. *Organizational Research Methods*, *18*(4), 679–703. https://doi.org/10.1177/1094428115574518

Zimmerman, D. W. (1975). Probability spaces, hilbert spaces, and the axioms of test theory. *Psychometrika*, *40*(3), 395–412. https://doi.org/10.1007/BF02291765

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, *4*(2), 223–233. https://doi.org/10.1080/15434300701375832

Appendix

Die empirischen Studien wurden beim Bundesministerium der Verteidigung (BMVg) Referat P III 5 unter der Registriernummer 2/02/17 genehmigt. Die militärische und zivile Gleichstellungsbeauftragte, Schwerbehindertenvertretung beim BMVg, der Hauptpersonalrat beim BMVg und der Gesamtvertrauenspersonenausschuss beim BMVg haben zugestimmt. Der Veröffentlichung der Dissertation wurde von BMVg P III 5 am 25.09.2020 zugestimmt.

Die Arbeit wurde durch eine Korrekturleserin gegen- und korrekturgelesen.

Hiermit bestätige ich, Ursa Katharina Johanna Nagler-Nitzschner, geb. Nagler, dass ich die vorliegende Arbeit nur mit den angegebenen Hilfsmitteln erstellt habe gemäß § 6 (3) der Promotionsordnung der Lebenswissenschaftlichen Fakultät. Die Dissertation oder Teile davon wurden nicht bereits bei einer anderen wissenschaftlichen Einrichtung eingereicht, angenommen oder abgelehnt. Zudem hat keine Zusammenarbeit mit gewerblichen Promotionsberatern stattgefunden. Ich habe die dem angestrebten Verfahren zugrunde liegende Promotionsordnung zur Kenntnis genommen und mich nicht nicht anderwärts um einen Doktorgrad beworben bzw. besitze keinen entsprechenden Doktorgrad. Die Grundsätze der Humboldt-Universität zu Berlin zur Sicherung guter wissenschaftlicher Praxis wurden eingehalten.

_____ _____

Ort, Datum Unterschrift

Detailed results Study 1

Working memory figural.

Item Set 1. In the first item set, the LRT (p = .23)[12] and the Martin-Löf-test (p = .95) were not significant for the RM. The χ^2/df of the LRT was 1.43. IFA was not significant as well (p = .41). Because of the missing items in the LRT, T_{11}-statistic was computed and as well not significant (p = .97). Both item parameters correlated highly (r = .95, p < .001).

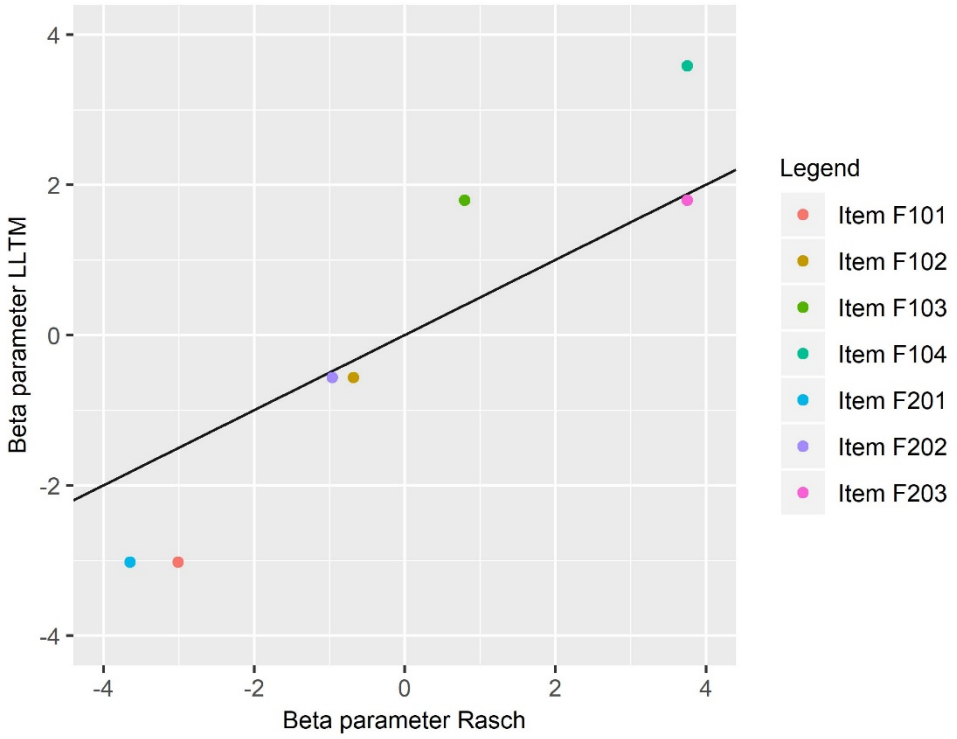


Figure 11. RM and LLTM beta parameter of Item Set 1.

[12] Items F203 and F104 had to be removed for the test because of missing response patterns within subgroups with split criterion mean.

Table 24

Item difficulty parameter for the RM and the LLTM - Item Set 1[13]

Item	Item difficulty parameter Rasch	Item difficulty parameter LLTM
F101	-3.010	-3.021
F201	-3.648	-3.021
F102	-0.682	-0.565
F202	-0.961	-0.565
F103	0.797	1.795
F203	3.752	1.795
F104	3.752	3.584

For the PCM, LRT showed no significance ($p = .98$)[14] as well as Martin-Löf-test ($p = .99$). IFA was not significant as well ($p = .22$). Item difficulty parameters of PCM and LPCM correlated with $r = .78$ ($p < .001$). Item difficulty parameters for the PCM and the LPCM without an a priori defined q-matrix correlated with $r = .92$ ($p < .001$).

[13] For a better comparison, four decimal places are shown here instead of the usual two.

[14] Items F102, F203 and F104 had to be removed for the test because of missing response patterns within subgroups with split criterion mean.

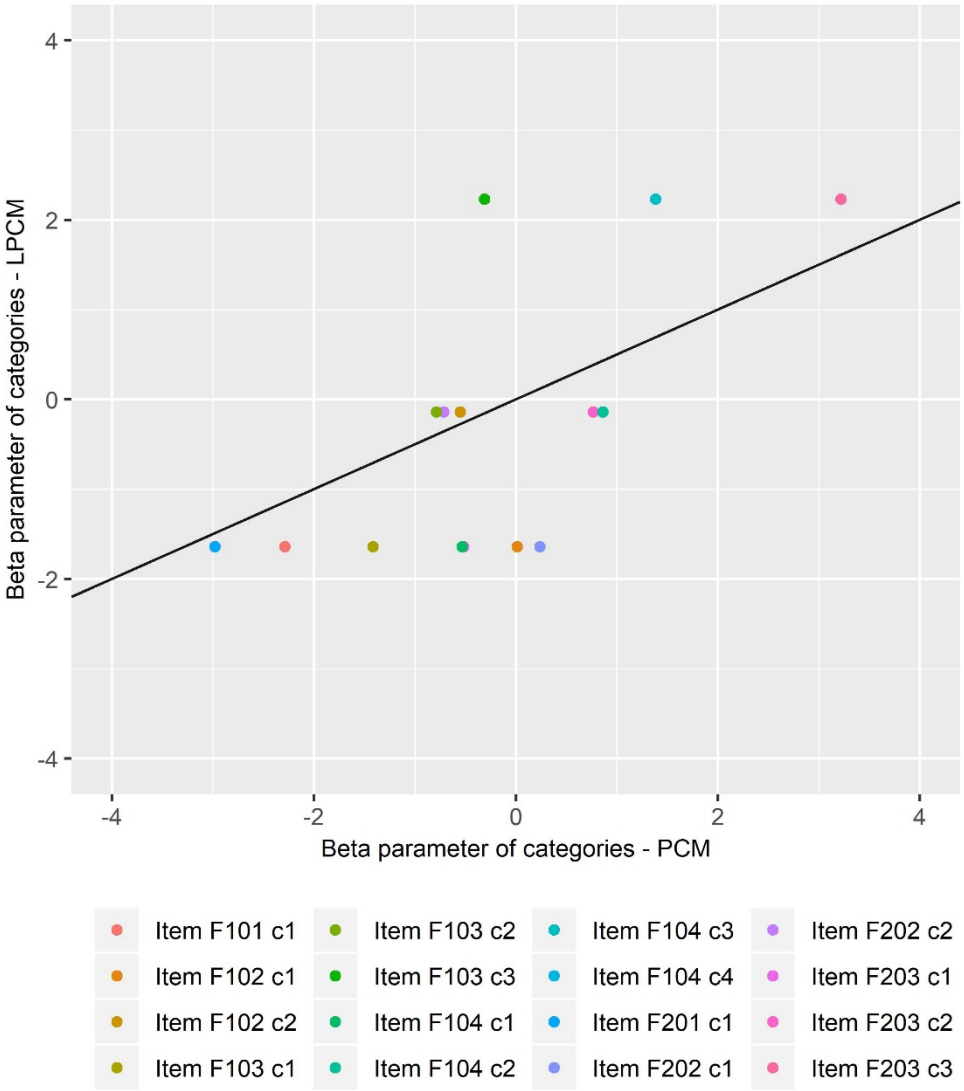


Figure 12. PCM and LPCM beta parameter of categories of Item Set 1.

In a simulation, random q-matrices with different ratios of 0's and 1's were generated and the LPCM calculated with those. The item difficulty parameter of the PCM and the newly calculated LPCM were correlated and the minimal correlation, the median, the mean, the 95th percentile and the maximum correlation determined as can be seen in Table 25. Missing values could not be calculated due to the properties of the artificially generated design matrix.

Table 25

Descriptive statistics for the correlations obtained from simulated weight matrices – Item Set 1[15]

$\%_1$	Min	Median	Mean	95%	Max
25	-.3388	.1617	.1695	.5235	.6403
30	-.4877	.1964	.1940	.5736	.7684
35	-.4576	.1269	.1427	.6063	.8309
40	-.4228	.2285	.2268	.6741	.7942
45	-.3879	.1849	.1881	.6122	.7593
50	-.3991	.2383	.2228	.5934	.6708
55	-.4469	.2658	.2355	.5966	.6975
60	-.2609	.2504	.2424	.5889	.7096
65	-.2351	.2823	.2716	.6421	.7210
70	-.3707	.3117	.2849	.6277	.7438

Note. $\%_1$ = occupancy with 1's in the weight matrix; Min = minimum correlation, 95% = 95th percentile; Max = maximum correlation.

In a second simulation, q-matrices were permutated. The results of the descriptive statistics can be seen in Table 26.

Table 26

Descriptive statistics for the correlations obtained from permutated simulated weight matrices – Item Set 1

Min	Median	Mean	95%	Max
.2524	.7357	.7005	.8223	.8223

Note. Min = minimum correlation, 95% = 95th percentile; Max = maximum correlation.

[15] Based on Baghaei and Hohensinn (2017), four decimal places are reported for all simulations.

Table 27

Item difficulty parameter for the PCM and the LPCM - Item Set 1

Item	Category	Item difficulty parameter PCM	Item difficulty parameter LPCM
F101	1	-2.288	-1.639
F201	1	-2.981	-1.639
F102	1	0.013	-1.639
F102	2	-0.552	-0.138
F202	1	0.236	-1.639
F202	2	-0.713	-0.138
F103	1	-1.414	-1.639
F103	2	-0.787	-0.138
F103	3	-0.311	2.231
F203	1	-0.516	-1.639
F203	2	0.765	-0.138
F203	3	3.218	2.231
F104	1	-0.530	-1.639
F104	2	0.860	-0.138
F104	3	1.383	2.231
F104	4	3.616	5.466

Item Set 2. In the second item set, the LRT ($p = .88$)[16] and the Martin-Löf-test ($p = .73$) were not significant for the RM. The χ^2/df of the LRT was 0.29. The T_{11}-statistic showed no significant result either ($p = .72$). IFA was not significant as well ($p = .24$). Difficulty parameters of the RM and the LLTM correlated highly with $r = .94$ ($p < .001$).

[16] Items 301, F101 and F104 had to be removed for the test because of missing response patterns within subgroups with split criterion mean.

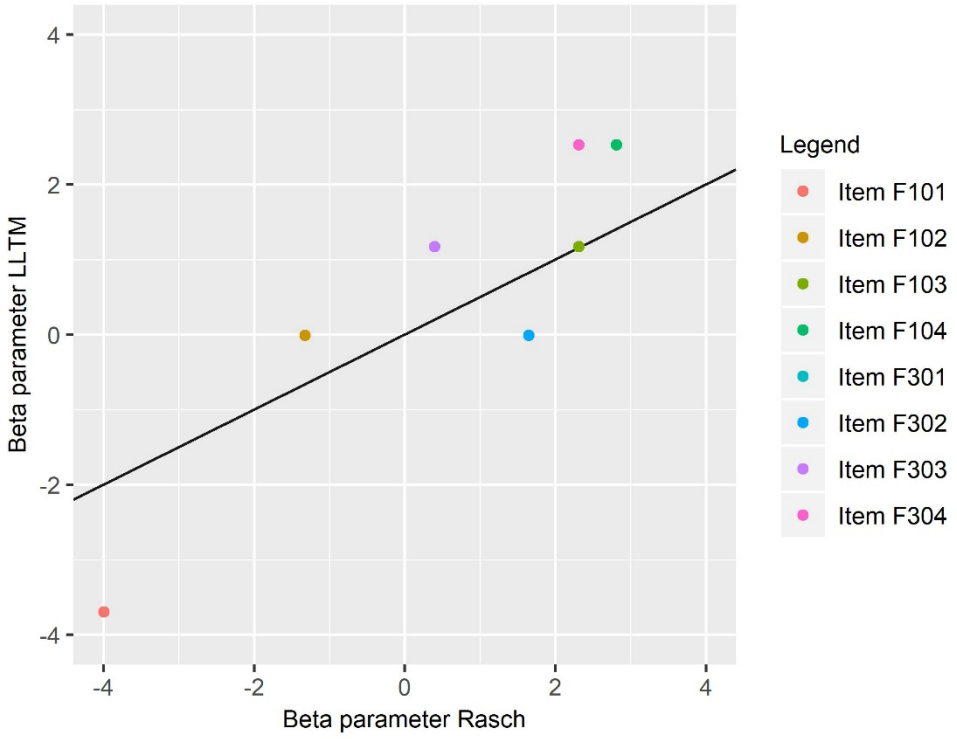


Figure 13. RM and LLTM beta parameter of Item Set 2.

Table 28

Item difficulty parameter for the RM and the LLTM - Item Set 2

Item	Item difficulty parameter Rasch	Item difficulty parameter LLTM
F301	-4.161	-3.695
F101	-3.997	-3.695
F302	1.649	-0.009
F102	-1.326	-0.009
F303	0.398	1.173
F103	2.312	1.173
F304	2.312	2.531
F104	2.813	2.531

For the PCM, LRT showed no significance (p = .84)[17] as well as Martin-Löf-test (p = .99). IFA was not significant as well (p = .21). Item difficulty parameters of PCM and LPCM correlated with r = .73 (p < .001). Item difficulty parameters for the PCM and the LPCM without an a priori defined q-matrix correlated with r = .88 (p < .001).

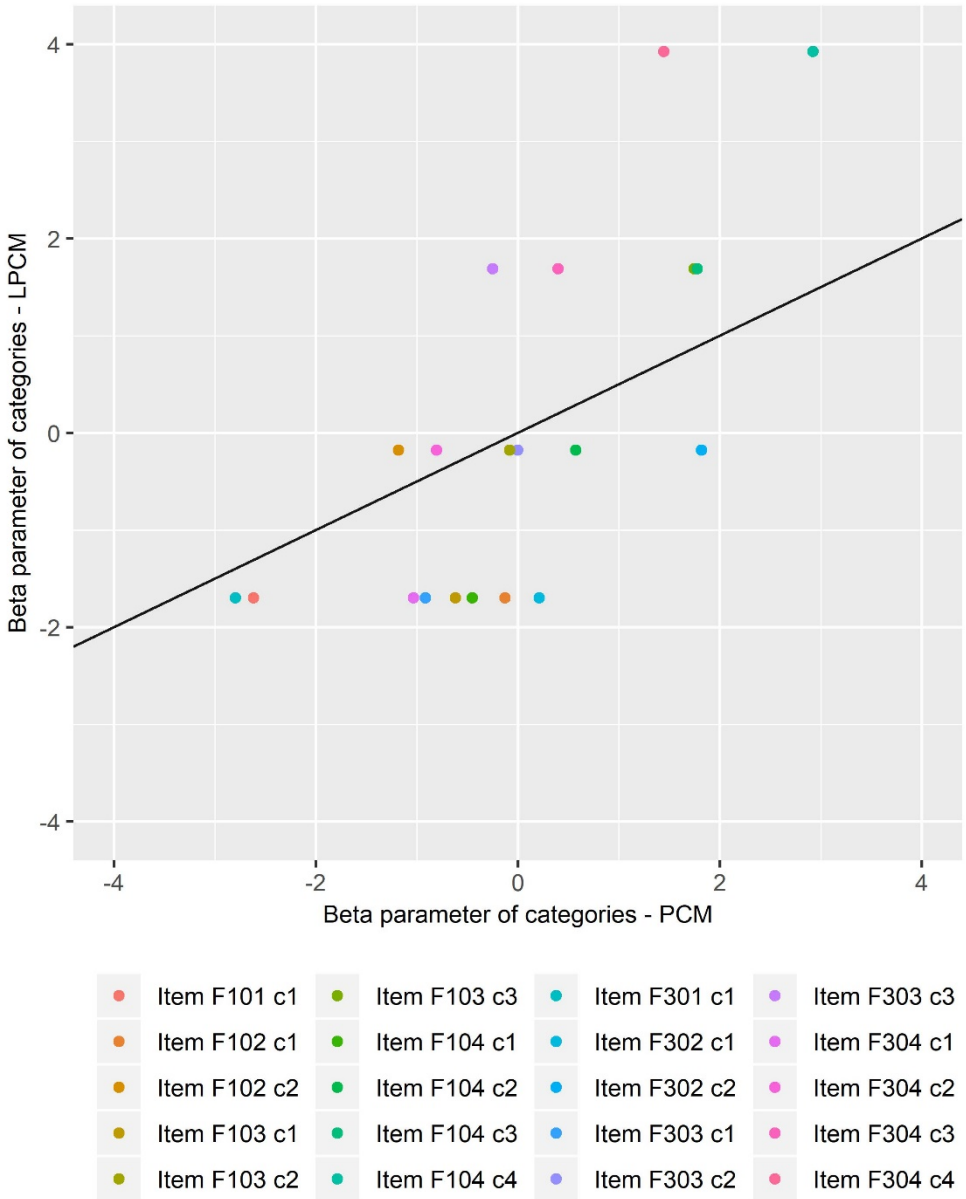


Figure 14. PCM and LPCM beta parameter of categories of Item Set 2.

[17] Items F301, F304 and F104 had to be removed for the test because of missing response patterns within subgroups with split criterion mean.

In a simulation, random q-matrices with different ratios of 0's and 1's were generated and the LPCM calculated with those. The item difficulty parameter of the PCM and the newly calculated LPCM were correlated and the minimal correlation, the median, the mean, the 95th percentile and the maximum correlation determined as can be seen in Table 29.

Table 29

Descriptive statistics for the correlations obtained from simulated weight matrices – Item Set 2

$\%_1$	Min	Median	Mean	95%	Max
20	-.4102	.1591	.1653	.5049	.6751
25	-.4398	.1652	.1537	.5256	.7726
30	-.3131	.1507	.1700	.5158	.7480
35	-.4032	.2371	.2143	.5487	.6824
40	-.2769	.1934	.1913	.5236	.6091
45	-.3338	.1963	.2024	.5395	.6379
50	-.3556	.1964	.1915	.5447	.6651
55	-.3717	.2357	.2300	.5397	.6761
60	-.2486	.2800	.2701	.6066	.7035
65	-.3470	.2357	.2370	.5776	.6518
70	-.2881	.2501	.2534	.5695	.7911

Note. $\%_1$ = occupancy with 1's in the weight matrix; Min = minimum correlation, 95% = 95th percentile; Max = maximum correlation.

In a second simulation, q-matrices were permutated. The results of the descriptive statistics can be seen in Table 30.

Table 30

Descriptive statistics for the correlations obtained from permutated simulated weight matrices – Item Set 2

Min	Median	Mean	95%	Max
.3110	.6615	.6026	.7099	.7322

Note. Min = minimum correlation, 95% = 95th percentile; Max = maximum correlation.

Table 31

Item difficulty parameter for the PCM and the LPCM - Item Set 2

Item	Category	Item difficulty parameter PCM	Item difficulty parameter LPCM
F301	1	-2.797	-1.695
F101	1	-2.618	-1.695
F302	1	0.212	-1.695
F302	2	1.819	-0.175
F102	1	-0.129	-1.695
F102	2	-1.184	-0.175
F303	1	-0.918	-1.695
F303	2	-0.001	-0.175
F303	3	-0.251	1.689
F103	1	-0.620	-1.695
F103	2	-0.081	-0.175
F103	3	1.746	1.689
F304	1	-1.034	-1.695
F304	2	-0.808	-0.175
F304	3	0.399	1.689
F304	4	1.446	3.928
F104	1	-0.451	-1.695
F104	2	0.574	-0.175
F104	3	1.775	1.689
F104	4	2.919	3.928

Item Set 3. Due to missing response patterns of item F201 and F203, both items had to be excluded. There were no sufficient data left to calculate a LRT or Martin-Löf-test. However, the T_{11}-statitic was not significant ($p = .21$) as well as the IFA ($p = .98$). The item difficulty parameter of the RM and the LLTM correlated highly with $r = .73$ ($p = .17$), although not significant.

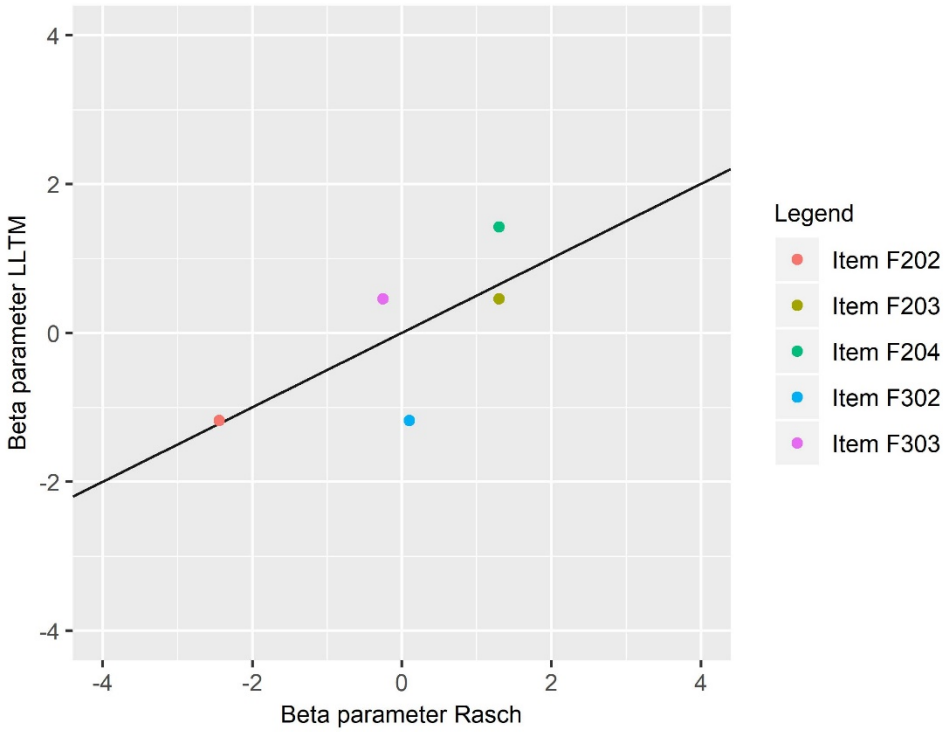


Figure 15. RM and LLTM beta parameter of Item Set 3.

Table 32

Item difficulty parameter for the RM and the LLTM - Item Set 3

Item	Item difficulty parameter Rasch	Item difficulty parameter LLTM
F201	---	---
F301	---	---
F202	-2.446	-1.174
F302	0.098	-1.174
F203	1.300	0.460
F303	-0.253	0.460
F204	1.300	1.428

For the PCM, LRT could not be performed because of too few responses. Martin-Löf-test ($p = .96$) was not significant. IFA was not significant as well ($p = .54$). Item difficulty parameters of PCM and LPCM correlated with $r = .80$ ($p < .001$). Item difficulty parameters for the PCM and the LPCM without an a priori defined q-matrix correlated with $r = .93$ ($p < .001$).

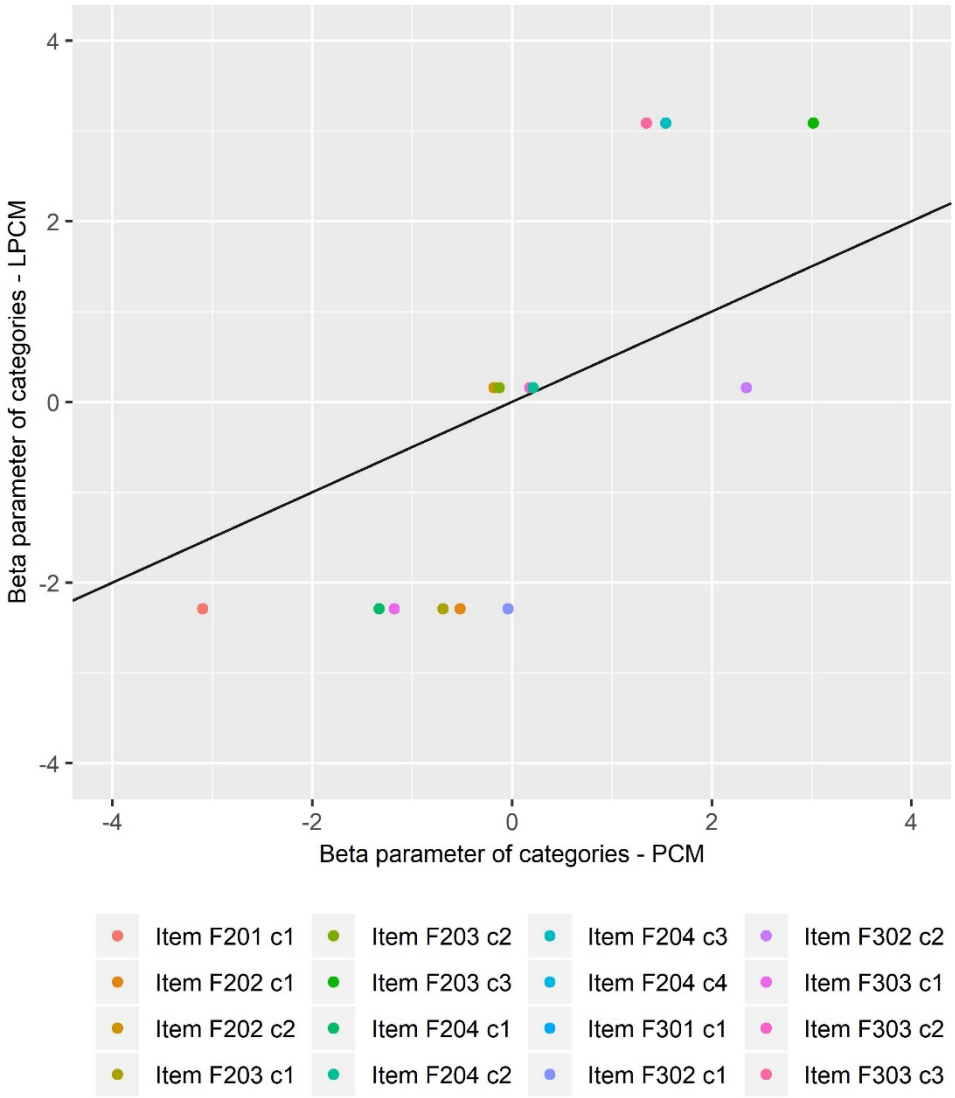


Figure 16. PCM and LPCM beta parameter of categories of Item Set 3.

In a simulation, random q-matrices with different ratios of 0's and 1's were generated and the LPCM calculated with those. The item difficulty parameter of the PCM and the newly calculated LPCM were correlated and the minimal correlation, the median, the mean, the 95th percentile and the maximum correlation determined as can be seen in Table 33. Missing values could not be calculated due to the properties of the artificially generated design matrix.

Table 33

Descriptive statistics for the correlations obtained from simulated weight matrices – Item Set 3

$\%_1$	Min	Median	Mean	95%	Max
20	-.7112	.1682	.1826	.5560	.8528
25	-.4171	.1950	.1990	.5364	.6986
30	-.2327	.2208	.2353	.5985	.7395
35	-.3034	.2081	.2011	.5521	.6862
40	-.4462	.2760	.2317	.5564	.6984
45	-.4818	.2369	.2251	.5382	.7080
50	-.3280	.2292	.2199	.5852	.6681
55	-.2926	.2228	.2133	.5842	.6684
60	-.3835	.3037	.2916	.6511	.8014
65	-.2039	.2887	.2880	.6650	.7128

Note. $\%_1$ = occupancy with 1's in the weight matrix; Min = minimum correlation, 95% = 95th percentile; Max = maximum correlation.

In a second simulation, q-matrices were permutated. The results of the descriptive statistics can be seen in Table 34.

Table 34

Descriptive statistics for the correlations obtained from permutated simulated weight matrices – Item Set 3

Min	Median	Mean	95%	Max
.2927	.6695	.6190	.7741	.7953

Note. Min = minimum correlation, 95% = 95th percentile; Max = maximum correlation.

Table 35

Item difficulty parameter for the PCM and the LPCM - Item Set 3

Item	Category	Item difficulty parameter PCM	Item difficulty parameter LPCM
F201	1	-3.097	-2.289
F301	1	-4.844	-2.289
F202	1	-0.520	-2.289
F202	2	-0.181	0.160
F302	1	-0.040	-2.289
F302	2	2.345	0.160
F203	1	-0.692	-2.289
F203	2	-0.129	0.160
F203	3	3.015	3.087
F303	1	-1.178	-2.289
F303	2	0.177	0.160
F303	3	1.346	3.087
F204	1	-1.331	-2.289
F204	2	0.209	0.160
F204	3	1.539	3.087
F204	4	3.382	5.961

Item Set 1 - 3. The RM with imputed data with $k = 5$ showed a not significant LRT ($p = .39$) as well as a Martin-Löf-test ($p = .93$). IFA was significant ($p < .05$). Item parameter of the RM and the LLTM correlated highly with $r = .95$ ($p < .001$).

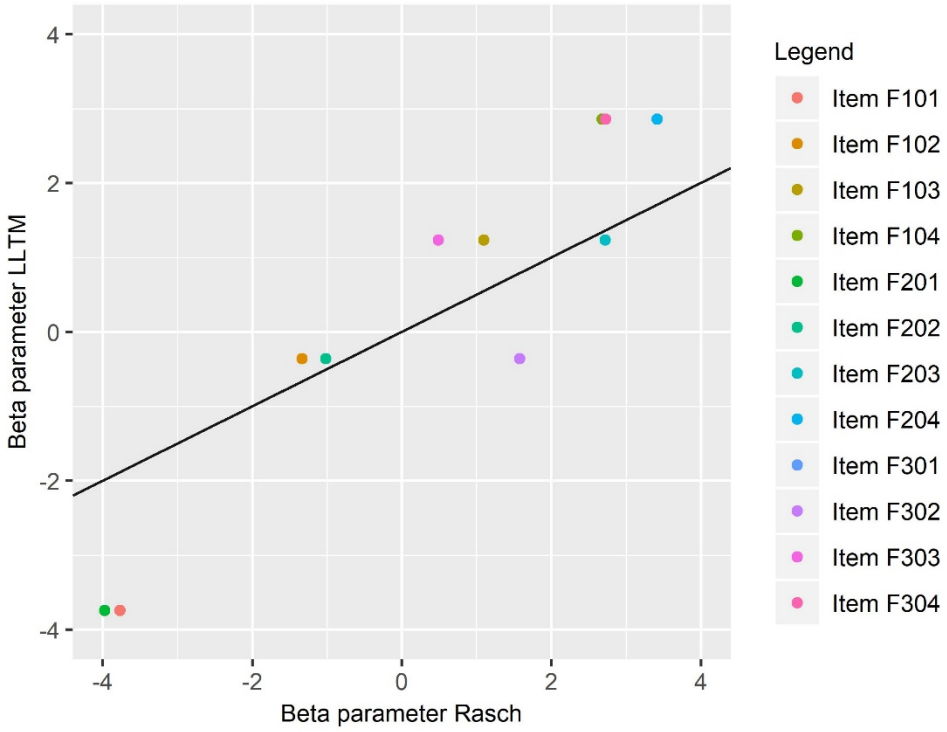


Figure 17. RM and LLTM beta parameter of Item Set 1 – 3.

Table 36

Item difficulty parameter for the RM and the LLTM - Item Set 1-3

Item	Item difficulty parameter Rasch	Item difficulty parameter LLTM
F101	-3.772	-3.741
F201	-3.975	-3.741
F301	-4.594	-3.741
F102	-1.334	-0.358
F202	-1.021	-0.358
F302	1.578	-0.358
F103	1.098	1.237
F203	2.720	1.237
F303	0.486	1.237
F104	2.677	2.863
F204	3.415	2.863
F304	2.722	2.863

For the PCM, LRT showed no significance ($p = .72$) as well as Martin-Löf-test ($p = .83$). Item difficulty parameters of PCM and LPCM correlated with $r = .81$ ($p < .001$). Item difficulty parameters for the PCM and the LPCM without an a priori defined q-matrix correlated with $r = 1$ ($p < .001$).

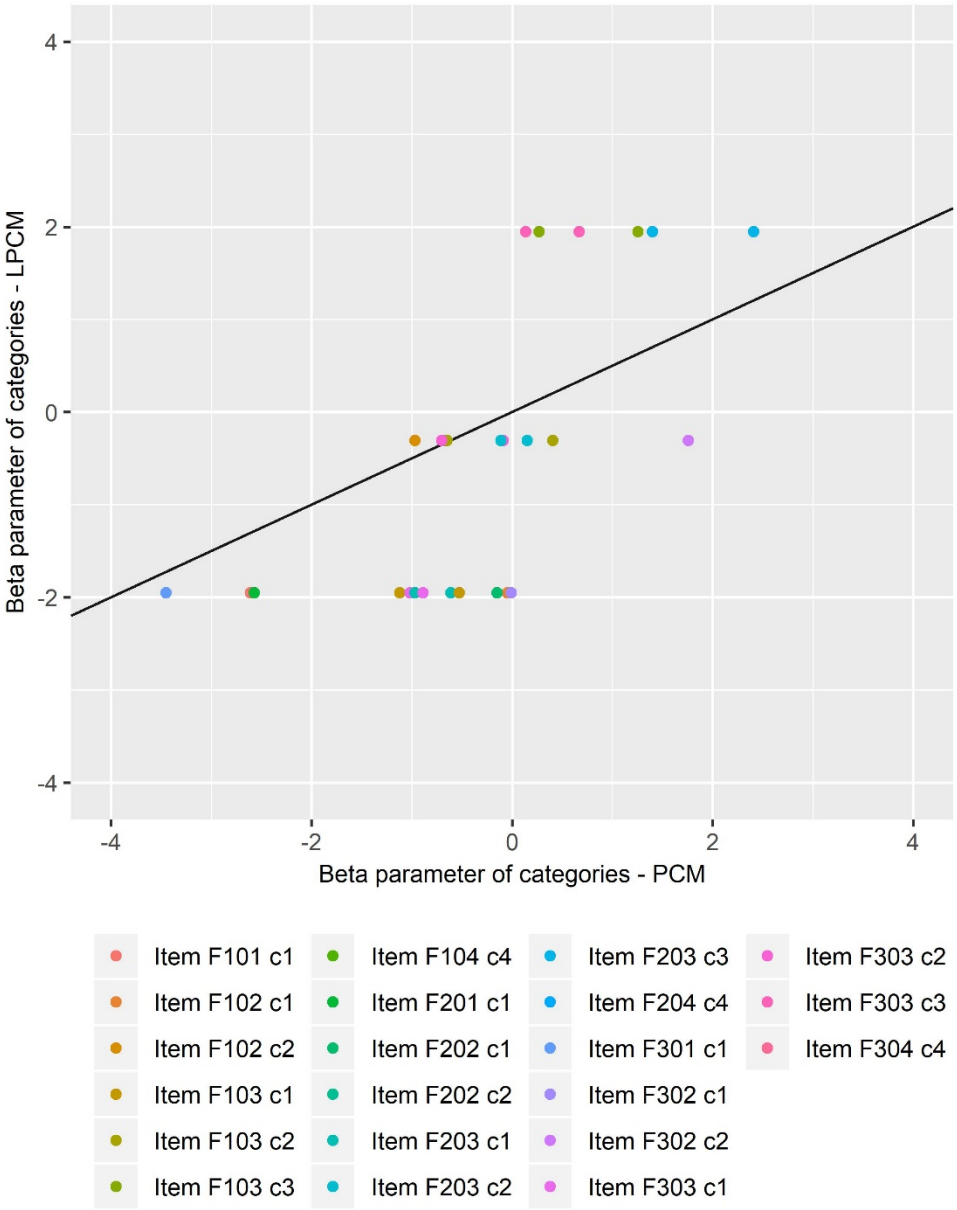


Figure 18. PCM and LPCM beta parameter of categories of Item Set 1 -3.

In a simulation, random q-matrices with different ratios of 0's and 1's were generated and the LPCM calculated with those. The item difficulty parameter of the PCM and the newly calculated LPCM were correlated and the minimal correlation, the median, the mean, the 95th percentile and the maximum correlation determined for each imputed dataset as can be seen in Table 37 to Table 38. Missing values could not be calculated due to the properties of the artificially generated design matrix.

Table 37

Descriptive statistics for the correlations obtained from simulated weight matrices for 65% occupancy – Item Set 1 - 3

$\%_1$	Min	Median	Mean	95%	Max
65	-.1932	.1562	.1681	.4376	.5179
65	-.2897	.1635	.1650	.4416	.5078
65	-.2627	.1508	.1483	.4545	.5958
65	-.2178	.1588	.1714	.4780	.6071
65	-.4525	.1898	.1769	.4604	.5753

Note. $\%_1$ = occupancy with 1's in the weight matrix; Min = minimum correlation, 95% = 95th percentile; Max = maximum correlation.

Table 38

Descriptive statistics for the correlations obtained from simulated weight matrices for 70% occupancy – Item Set 1 - 3

$\%_1$	Min	Median	Mean	95%	Max
70	-.2229	.1507	.1535	.4007	.6007
70	-.2089	.1667	.1419	.4508	.5563
70	-.1995	.1557	.1724	.4691	.5344
70	-.2854	.1861	.1648	.4381	.6025
70	-.3678	.1570	.1584	.4308	.5846

Note. $\%_1$ = occupancy with 1's in the weight matrix; Min = minimum correlation, 95% = 95th percentile; Max = maximum correlation.

In a second simulation, q-matrices were permutated. The results of the descriptive statistics for each imputed dataset can be seen in Table 39.

Table 39

Descriptive statistics for the correlations obtained from permutated simulated weight matrices – Item Set 1 - 3

Min	Median	Mean	95%	Max
.2903	.7159	.6504	.8202	.8202
.3159	.7689	.6914	.8423	.8423
.3281	.7231	.6638	.8202	.8202
.3467	.6911	.6420	.7962	.7962
.3033	.7885	.7091	.8609	.8609

Note. Min = minimum correlation, 95% = 95th percentile; Max = maximum correlation.

Table 40

Item difficulty parameter for the PCM and the LPCM - Item Set 1 - 3

Item	Category	Item difficulty parameter PCM	Item difficulty parameter LPCM
F101	1	-2.609	-1.949
F201	1	-2.573	-1.949
F301	1	-3.451	-1.949
F101	1	-0.050	-1.949
F102	2	-0.968	-0.308
F201	1	-0.152	-1.949
F202	2	-0.675	-0.308
F301	1	-0.009	-1.949
F302	2	1.756	-0.308
F101	1	-1.122	-1.949
F102	2	-0.649	-0.308
F103	3	0.269	1.948
F201	1	-0.613	-1.949
F202	2	0.148	-0.308

continued

continued

Item	Category	Item difficulty parameter PCM	Item difficulty parameter LPCM
F203	3	2.407	1.948
F301	1	-1.018	-1.949
F302	2	-0.091	-0.308
F303	3	0.133	1.948
F101	1	-0.526	-1.949
F102	2	0.405	-0.308
F103	3	1.253	1.948
F104	4	2.757	4.826
F201	1	-0.970	-1.949
F202	2	-0.114	-0.308
F203	3	1.400	1.948
F204	4	3.779	4.826
F301	1	-0.891	-1.949
F302	2	-0.703	-0.308
F303	3	0.666	1.948
F304	4	2.212	4.826

Item Set 4. Too many items had to be removed due to missing responses in subgroups (split criterion mean), hence, no p-value for the LRT is reported. The Martin-Löf-test was not significant ($p = 1.0$), however, the T_{11}-statistic was significant, showing a dependency between item F202 and F103. $T_{1\ell}$-statistic shows too similar response patterns between items. T_{10}-statistic was significant as well. IFA was not significant ($p = .08$). Although not all items were locally independent, the RM and LLTM item difficulty parameters were calculated and correlated with $r = .97$ ($p < .001$).

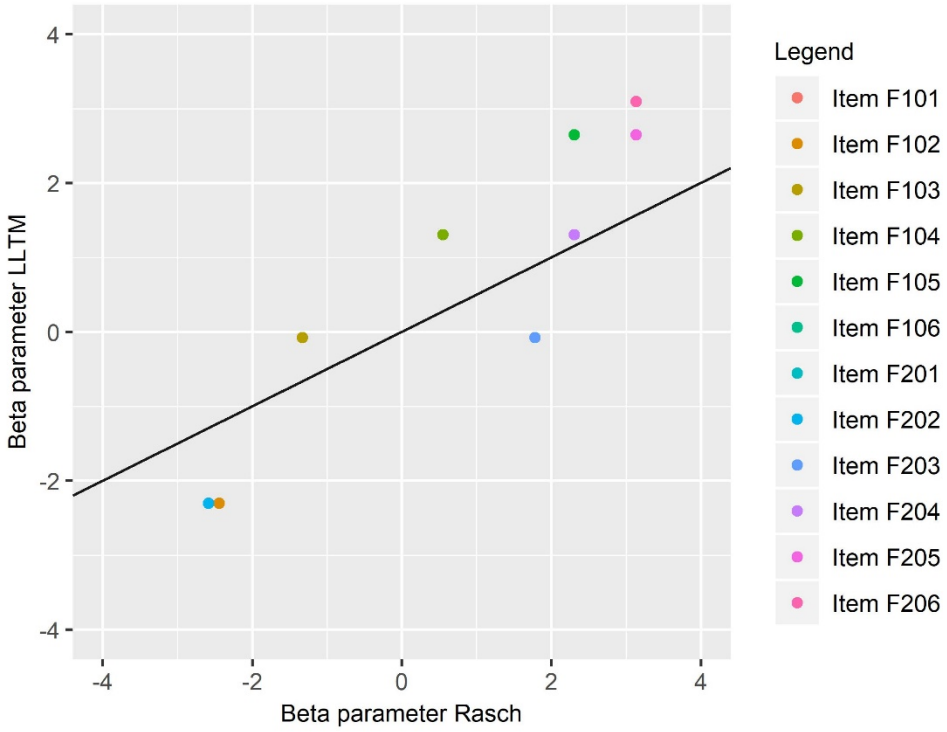


Figure 19. RM and LLTM beta parameter of Item Set 4.

Table 41

Item difficulty parameter for the RM and the LLTM - Item Set 4

Item	Item difficulty parameter Rasch	Item difficulty parameter LLTM
F101	-4.610	-4.683
F201	-5.371	-4.683
F102	-2.446	-2.300
F202	-2.588	-2.300
F103	-1.329	-0.073
F203	1.780	-0.073
F104	0.550	1.307
F204	2.308	1.307
F105	2.308	2.650
F205	3.133	2.650
F106	3.133	3.099
F206	3.133	3.099

For the PCM, LRT could not be performed because of too few response patterns. Martin-Löf-test ($p = 1.0$) was not significant. However, IFA was significant ($p = .02$). Item difficulty parameters of PCM and LPCM correlated with $r = .89$ ($p < .001$). Item difficulty parameters for the PCM and the LPCM without an a priori defined q-matrix correlated with $r = .95$ ($p < .001$).

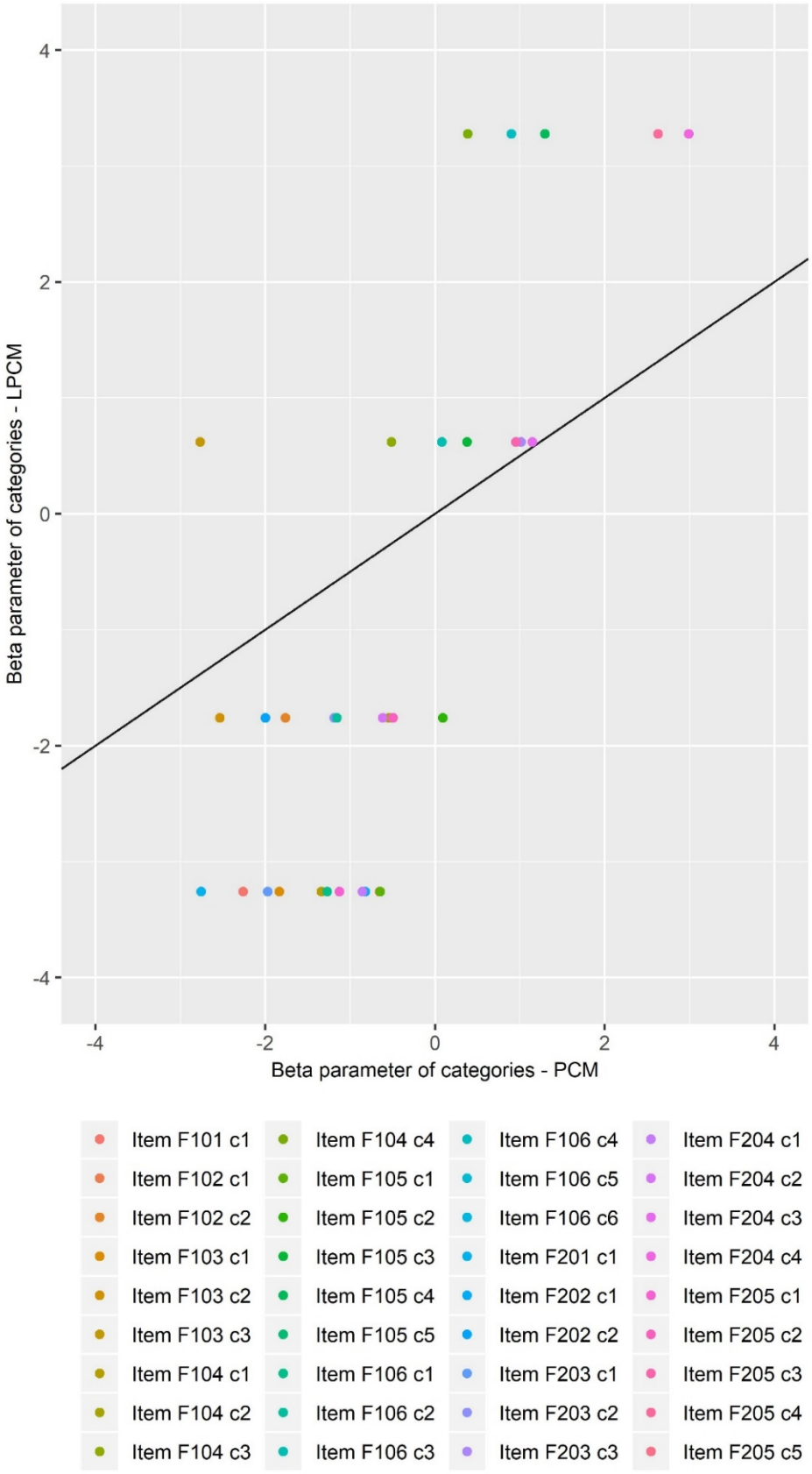


Figure 20. PCM and LPCM beta parameter of categories of Item Set 4.

In a simulation, random q-matrices with different ratios of 0's and 1's were generated and the LPCM calculated with those. The item difficulty parameter of the PCM and the newly calculated LPCM were correlated and the minimal correlation, the median, the mean, the 95th percentile and the maximum correlation determined as can be seen in Table 42. Missing values could not be calculated due to the properties of the artificially generated design matrix.

Table 42

Descriptive statistics for the correlations obtained from simulated weight matrices – Item Set 4

$\%_1$	Min	Median	Mean	95%	Max
20	-.2518	.1135	.1278	.4423	.4822
25	-.4050	.1252	.1242	.4206	.6134
30	-.2757	.0953	.1123	.3962	.4949
35	-.2662	.1275	.1404	.4267	.5927
40	-.3029	.0863	.0985	.4017	.4876
45	-.2208	.0969	.1300	.4209	.4836
50	-.2547	.1577	.1561	.3762	.5483
55	-.1912	.1435	.1446	.3642	.5714
60	-.3249	.0996	.1208	.4121	.5763
65	-.2520	.1370	.1402	.3877	.5696

Note. $\%_1$ = occupancy with 1's in the weight matrix; Min = minimum correlation, 95% = 95th percentile; Max = maximum correlation.

In a second simulation, q-matrices were permutated. The results of the descriptive statistics can be seen in Table 43.

Table 43

Descriptive statistics for the correlations obtained from permutated simulated weight matrices – Item Set 4

Min	Median	Mean	95%	Max
.5240	.8573	.8168	.9112	.9140

Note. Min = minimum correlation, 95% = 95th percentile; Max = maximum correlation.

Table 44

Item difficulty parameter for the PCM and the LPCM - Item Set 4

Item	Category	Item difficulty parameter PCM	Item difficulty parameter LPCM
F101	1	-2.261	-3.258
F201	1	-2.756	-3.258
F102	1	-0.653	-3.258
F102	2	-1.761	-1.759
F202	1	-0.821	-3.258
F202	2	-1.996	-1.759
F103	1	-1.833	-3.258
F103	2	-2.533	-1.759
F103	3	-2.765	0.621
F203	1	-1.970	-3.258
F203	2	-1.185	-1.759
F203	3	1.015	0.621
F104	1	-1.340	-3.258
F104	2	-0.544	-1.759
F104	3	-0.511	0.621
F104	4	0.385	3.278
F204	1	-0.854	-3.258
F204	2	-0.616	-1.759

continued

continued

Item	Category	Item difficulty parameter PCM	Item difficulty parameter LPCM
F204	3	1.148	0.621
F204	4	2.990	3.278
F105	1	-0.648	-3.258
F105	2	0.091	-1.759
F105	3	0.378	0.621
F105	4	1.298	3.278
F105	5	3.408	6.927
F205	1	-1.127	-3.258
F205	2	-0.494	-1.759
F205	3	0.951	0.621
F205	4	2.629	3.278
F205	5	4.976	6.927
F106	1	-1.271	-3.258
F106	2	-1.155	-1.759
F106	3	0.082	0.621
F106	4	0.901	3.278
F106	5	3.521	6.927
F106	6	5.322	10.149

Item Set 5. The LRT was not significant ($p = .48$)[18] as well as the Martin-Löf-test ($p = .99$). The χ^2/df of the LRT was 0.89. IFA was not significant ($p = .18$). The item difficulty parameter of RM and LLTM correlated ($r = .97$, $p < .001$).

[18] Items F103, F304, F305, F105, F306 and F106 had to be removed for the test because of missing response patterns within subgroups with split criterion mean.

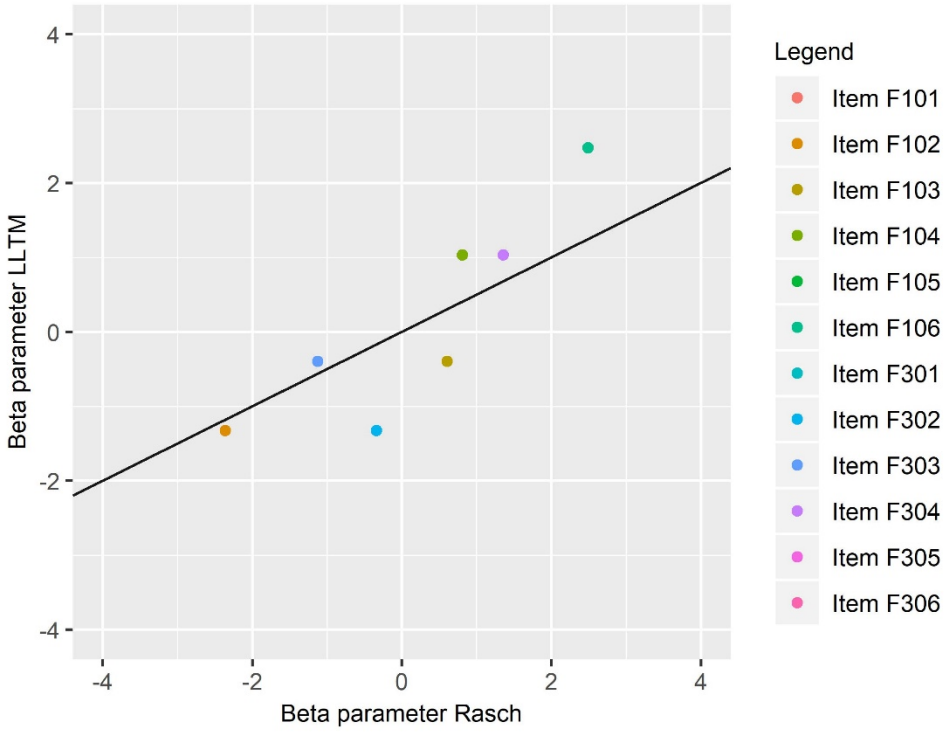


Figure 21. RM and LLTM beta parameter of Item Set 5.

Table 45

Item difficulty parameter for the RM and the LLTM - Item Set 5

Item	Item difficulty parameter Rasch	Item difficulty parameter LLTM
F301	-4.454	-4.270
F101	-4.454	-4.270
F302	-0.337	-1.323
F102	-2.362	-1.323
F303	-1.126	-0.392
F103	0.605	-0.392
F304	1.357	1.035
F104	0.808	1.035
F305	2.491	2.474
F105	2.491	2.474
F306	2.491	2.474
F106	2.491	2.474

For the PCM, LRT showed no significance ($p = .12$)[19] as well as Martin-Löf-test ($p = 1.0$). IFA was not significant as well ($p = .20$). Item difficulty parameters of PCM and LPCM correlated with $r = .82$ ($p < .001$). Item difficulty parameters for the PCM and the LPCM without an a priori defined q-matrix correlated with $r = .89$ ($p < .001$).

[19] Items F303, F304, F305, F105, F306 and F106 had to be removed for the test because of missing response patterns within subgroups with split criterion mean.

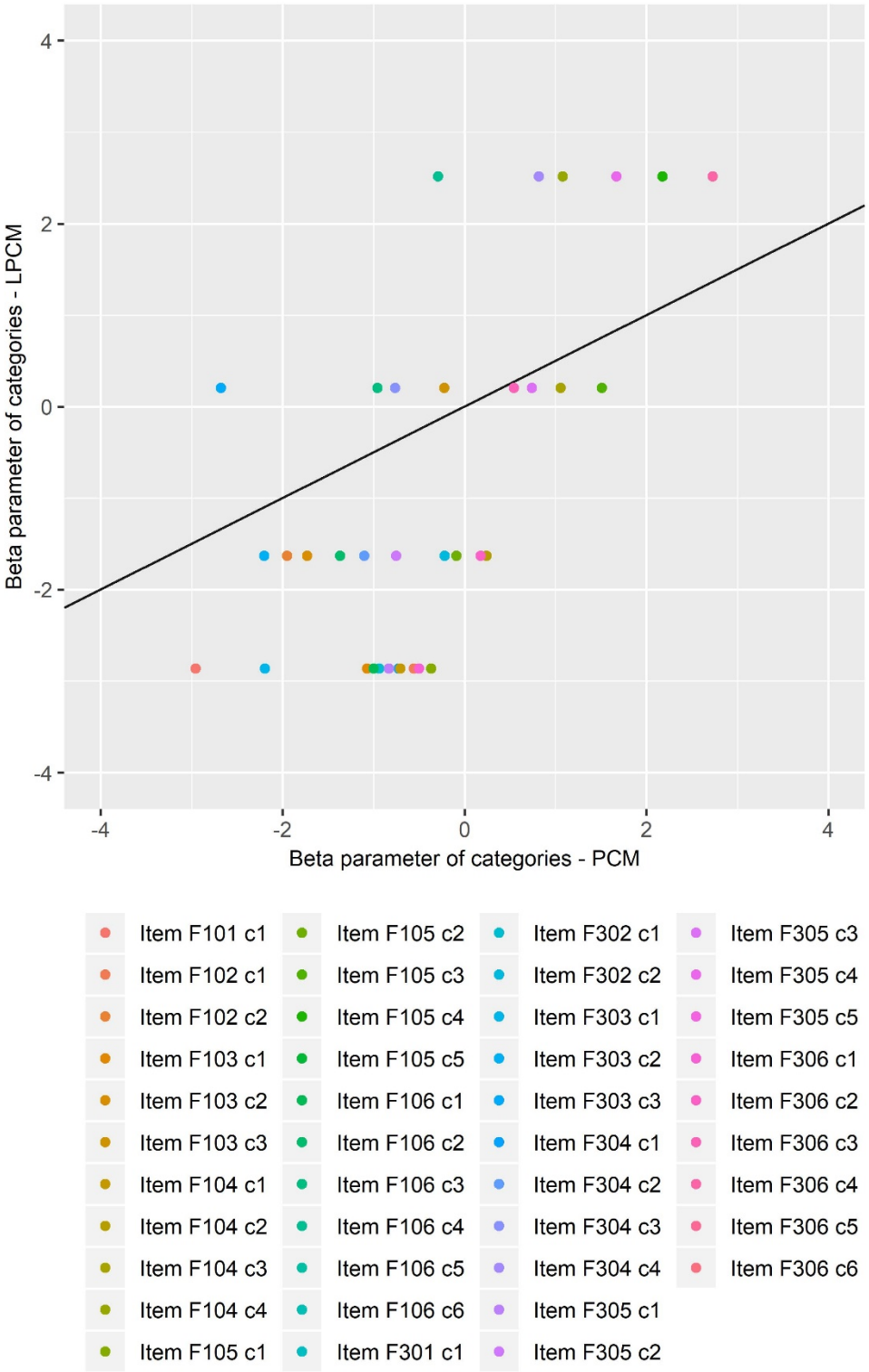


Figure 22. PCM and LPCM beta parameter of categories of Item Set 5.

In a simulation, random q-matrices with different ratios of 0's and 1's were generated and the LPCM calculated with those. The item difficulty parameter of the PCM

and the newly calculated LPCM were correlated and the minimal correlation, the median, the mean, the 95th percentile and the maximum correlation determined as can be seen in Table 46.

Table 46

Descriptive statistics for the correlations obtained from simulated weight matrices – Item Set 5

$\%_1$	Min	Median	Mean	95%	Max
20	-.2825	.1205	.1213	.3601	.4423
25	-.2487	.1277	.1109	.3371	.4962
30	-.2809	.1257	.1245	.3489	.4974
35	-.2012	.1037	.1078	.3391	.4324
40	-.3464	.1547	.1477	.4305	.4613
45	-.3305	.1052	.1013	.3596	.4239
50	-.2006	.1470	.1499	.4033	.4645
55	-.3220	.1257	.1295	.3501	.6024
60	-.1977	.1670	.1420	.3495	.4678
65	-.2712	.1647	.1603	.3767	.4827
70	-.1792	.1526	.1643	.3983	.5234

Note. $\%_1$ = occupancy with 1's in the weight matrix; Min = minimum correlation, 95% = 95th percentile; Max = maximum correlation.

In a second simulation, q-matrices were permutated. The results of the descriptive statistics can be seen in Table 47.

Table 47

Descriptive statistics for the correlations obtained from permutated simulated weight matrices – Item Set 5

Min	Median	Mean	95%	Max
.5421	.7612	.7430	.8314	.8333

Note. Min = minimum correlation, 95% = 95th percentile; Max = maximum correlation.

Table 48

Item difficulty parameter for the PCM and the LPCM - Item Set 5

Item	Category	Item difficulty parameter PCM	Item difficulty parameter LPCM
F301	1	-2.957	-2.863
F101	1	-2.957	-2.863
F302	1	-0.937	-2.863
F302	2	-0.218	-1.628
F102	1	-0.557	-2.863
F102	2	-1.951	-1.628
F303	1	-2.194	-2.863
F303	2	-2.201	-1.628
F303	3	-2.677	0.207
F103	1	-1.072	-2.863
F103	2	-1.731	-1.628
F103	3	-0.222	0.207
F304	1	-0.726	-2.863
F304	2	-1.104	-1.628
F304	3	-0.765	0.207
F304	4	0.817	2.517
F104	1	-0.707	-2.863
F104	2	0.243	-1.628

continued

continued

Item	Category	Item difficulty parameter PCM	Item difficulty parameter LPCM
F104	3	1.055	0.207
F104	4	1.079	2.517
F305	1	-0.830	-2.863
F305	2	-0.753	-1.628
F305	3	0.738	0.207
F305	4	1.670	2.517
F305	5	2.706	5.131
F105	1	-0.368	-2.863
F105	2	-0.090	-1.628
F105	3	1.509	0.207
F105	4	2.174	2.517
F105	5	3.227	5.131
F306	1	-0.501	-2.863
F306	2	0.178	-1.628
F306	3	0.541	0.207
F306	4	2.726	2.517
F306	5	3.071	5.131
F306	6	3.497	6.676
F106	1	-0.999	-2.863
F106	2	-1.368	-1.628
F106	3	-0.956	0.207
F106	4	-0.292	2.517
F106	5	1.770	5.131
F106	6	2.135	6.676

Item Set 6. The LRT was not significant (p = .37)[20] as was the Martin-Löf-test (p = 1.0). The χ^2/df of the LRT was 1.06. T_{11}-statistic was not significant (p = .06). IFA was not significant (p = .26). Both item difficulty parameters correlated (r = .98, p < .001).

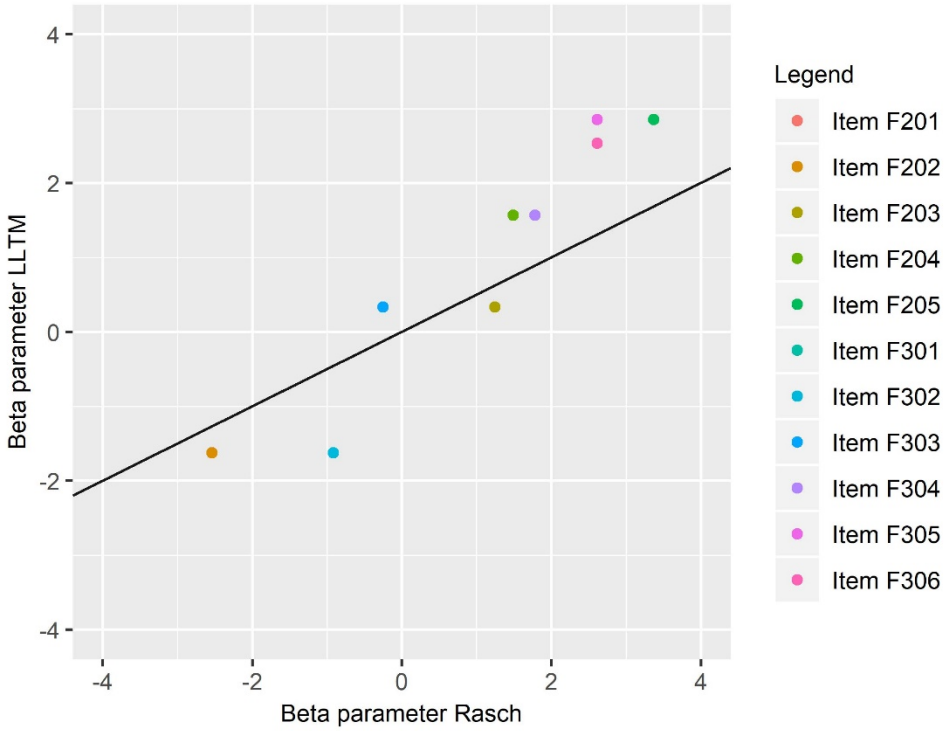


Figure 23. RM and LLTM beta parameter of Item Set 6.

[20] Items F203, F204, F304, F205, F305 and F306 had to be removed for the test because of missing response patterns within subgroups with split criterion mean.

Table 49

Item difficulty parameter for the RM and the LLTM - Item Set 6

Item	Item difficulty parameter Rasch	Item difficulty parameter LLTM
F201	-5.059	-4.403
F301	-4.330	-4.403
F202	-2.543	-1.622
F302	-0.916	-1.622
F203	1.242	0.334
F303	-0.251	0.334
F204	1.487	1.568
F304	1.779	1.568
F205	3.368	2.854
F305	2.611	2.854
F306	2.611	2.537

For the PCM, LRT showed no significance ($p = .50$)[21] as well as Martin-Löf-test ($p = 1.0$). IFA was significant ($p = .01$). Item difficulty parameters of PCM and LPCM correlated with $r = .88$ ($p < .001$). Item difficulty parameters for the PCM and the LPCM without an a priori defined q-matrix correlated with $r = .92$ ($p < .001$).

[21] Items F202, F203, F204, F304, F205, F305 and F306 had to be removed for the test because of missing response patterns within subgroups with split criterion mean.

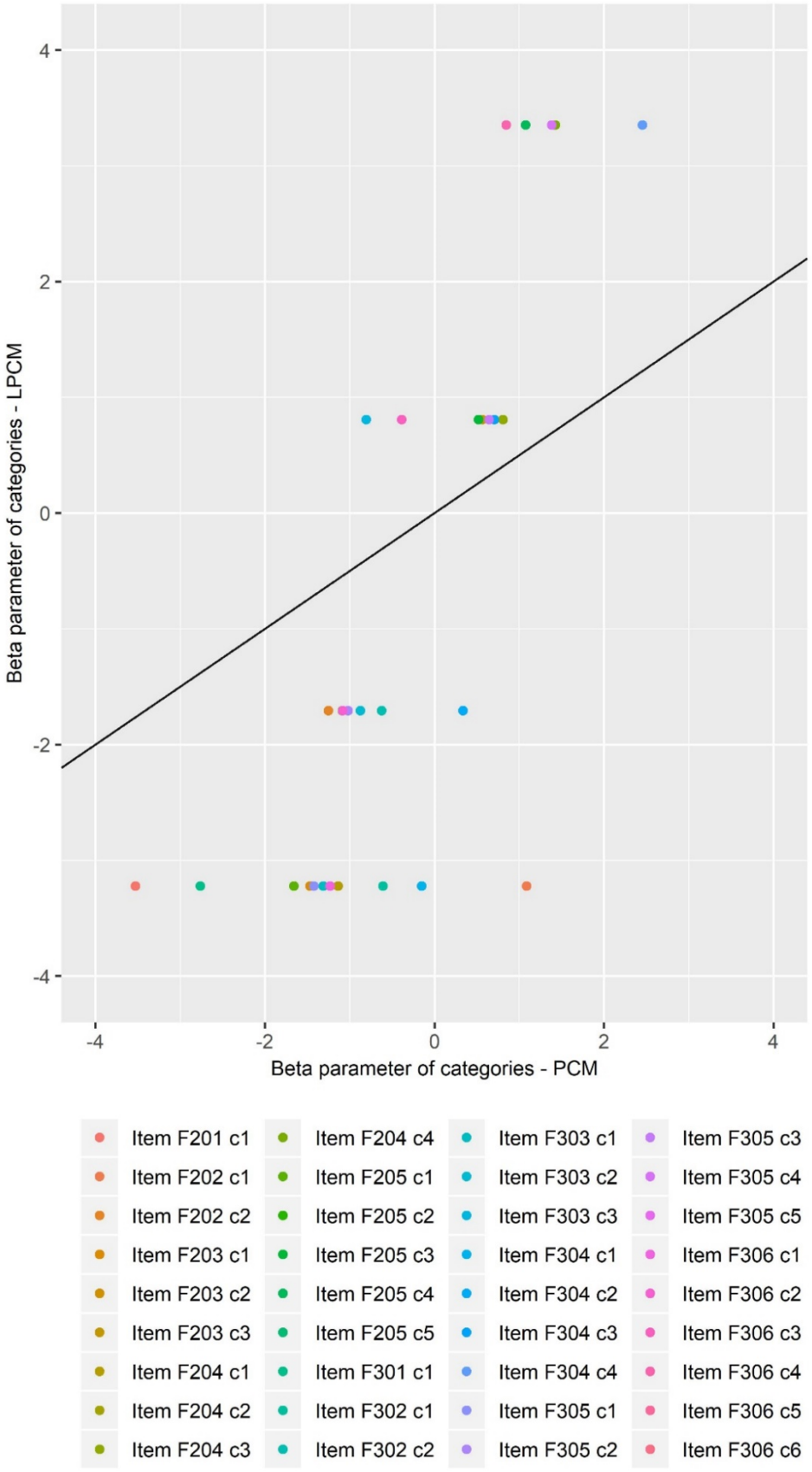


Figure 24. PCM and LPCM beta parameter of categories of Item Set 6.

In a simulation, random q-matrices with different ratios of 0's and 1's were generated and the LPCM calculated with those. The item difficulty parameter of the PCM and the newly calculated LPCM were correlated and the minimal correlation, the median, the mean, the 95th percentile and the maximum correlation determined as can be seen in Table 50.

Table 50

Descriptive statistics for the correlations obtained from simulated weight matrices – Item Set 6

$\%_1$	Min	Median	Mean	95%	Max
20	-.2972	.1229	.1222	.3904	.6040
25	-.4307	.1882	.1595	.4265	.5682
30	-.3641	.1390	.1164	.3820	.4867
35	-.1354	.1195	.1407	.4248	.5142
40	-.2188	.1429	.1385	.4033	.4918
45	-.3369	.1521	.1381	.3767	.4449
50	-.3199	.1955	.1805	.5058	.5843
55	-.2144	.1591	.1661	.4444	.5670
60	-.1751	.1837	.1651	.4039	.6126
65	-.3206	.1544	.1506	.4122	.4834
70	-.3640	.1506	.1325	.3961	.5195

Note. $\%_1$ = occupancy with 1's in the weight matrix; Min = minimum correlation, 95% = 95th percentile; Max = maximum correlation.

In a second simulation, q-matrices were permutated. The results of the descriptive statistics can be seen in Table 51.

Table 51

Descriptive statistics for the correlations obtained from permutated simulated weight matrices – Item Set 6

Min	Median	Mean	95%	Max
.4897	.8025	.7662	.8947	.8947

Note. Min = minimum correlation, 95% = 95th percentile; Max = maximum correlation.

Table 52

Item difficulty parameter for the PCM and the LPCM - Item Set 6

Item	Category	Item difficulty parameter PCM	Item difficulty parameter LPCM
F201	1	-3.528	-3.222
F301	1	-2.764	-3.222
F202	1	1.085	-3.222
F202	2	-1.250	-1.706
F302	1	-0.608	-3.222
F302	2	-0.624	-1.706
F203	1	-1.470	-3.222
F203	2	-1.020	-1.706
F203	3	0.566	0.808
F303	1	-1.313	-3.222
F303	2	-0.874	-1.706
F303	3	-0.807	0.808
F204	1	-1.138	-3.222
F204	2	-1.087	-1.706
F204	3	0.810	0.808
F204	4	1.427	3.351
F304	1	-0.152	-3.222
F304	2	0.338	-1.706

continued

continued

Item	Category	Item difficulty parameter PCM	Item difficulty parameter LPCM
F304	3	0.706	0.808
F304	4	2.452	3.351
F205	1	-1.657	-3.222
F205	2	-1.086	-1.706
F205	3	0.519	0.808
F205	4	1.074	3.351
F205	5	3.879	6.471
F305	1	-1.421	-3.222
F305	2	-1.020	-1.706
F305	3	0.644	0.808
F305	4	1.383	3.351
F305	5	3.233	6.471
F306	1	-1.233	-3.222
F306	2	-1.084	-1.706
F306	3	-0.387	0.808
F306	4	0.847	3.351
F306	5	1.961	6.471
F306	6	3.599	8.973

Item Set 4 - 6. To retrieve the missing data in Item Set 4 to 5, data was imputed with $k = 5$. Neither an LRT could be performed due to missing response patterns nor a Martin-Löf-test because of missing data. IFA was significant ($p < .05$). However, item parameter correlated with $r = .96$ ($p < .001$).

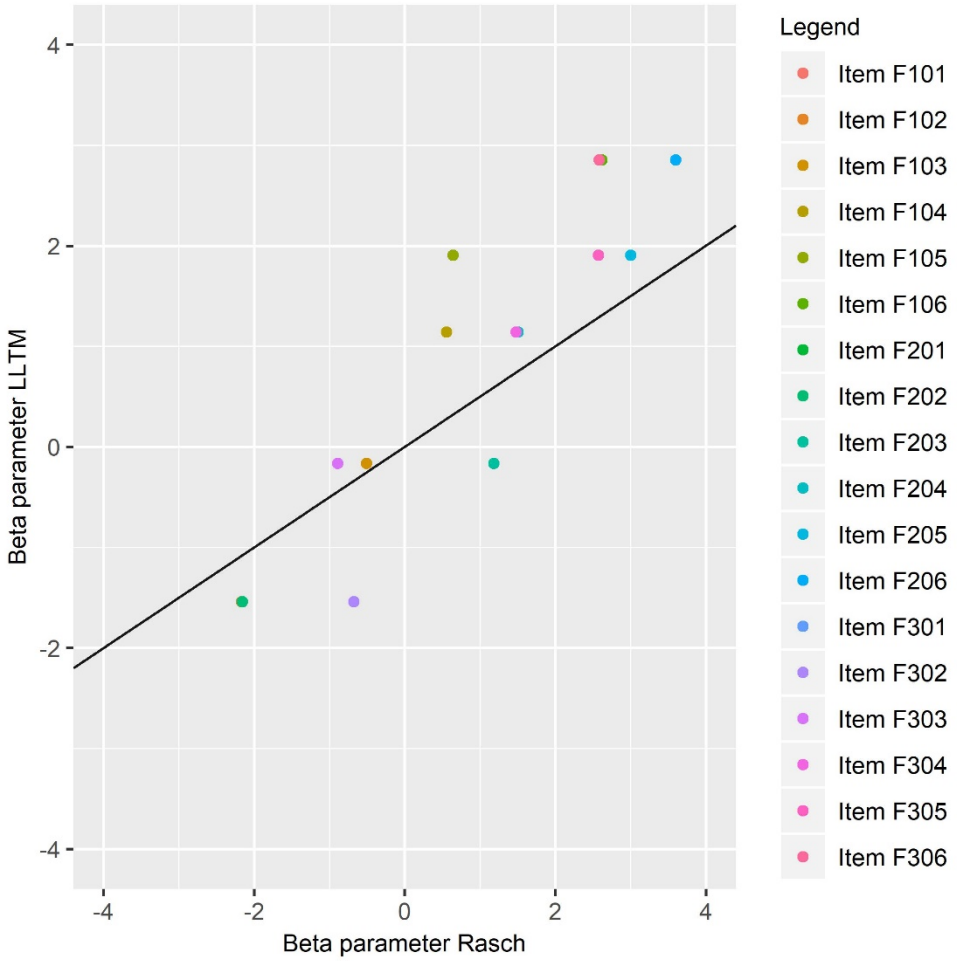


Figure 25. RM and LLTM beta parameter of Item Set 4 - 6.

Table 53

Item difficulty parameter for the RM and the LLTM - Item Set 4-6

Item	Item difficulty parameter Rasch	Item difficulty parameter LLTM
F101	-4.138	-4.202
F201	-4.890	-4.202
F301	-4.301	-4.202
F102	-2.167	-1.540
F202	-2.156	-1.540
F302	-0.677	-1.540
F103	-0.509	-0.164
F203	1.183	-0.164
F303	-0.892	-0.164
F104	0.554	1.141
F204	1.506	1.141
F304	1.477	1.141
F105	0.640	1.908
F205	3.001	1.908
F305	2.569	1.908
F106	2.619	2.856
F206	3.596	2.856
F306	2.583	2.856

For the PCM, LRT showed no significance ($p = .37$) as well as Martin-Löf-test ($p = 1.0$). Item difficulty parameters of PCM and LPCM correlated with $r = .88$ ($p < .001$). Item difficulty parameters for the PCM and the LPCM without an a priori defined q-matrix correlated with $r = 1$ ($p < .001$).

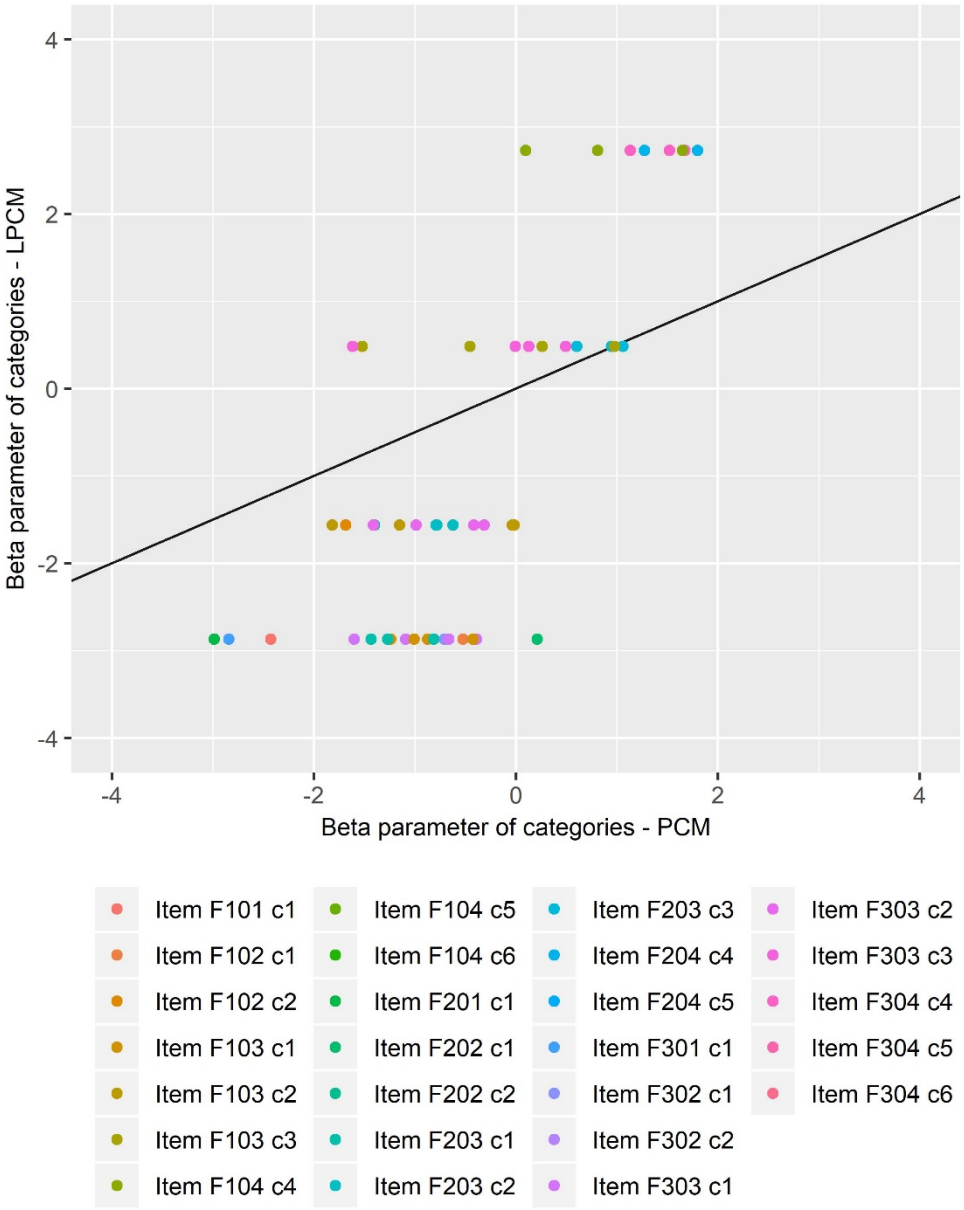


Figure 26. PCM and LPCM beta parameter of categories of Item Set 4 – 6.

In a simulation, random q-matrices with different ratios of 0's and 1's were generated and the LPCM calculated with those. The item difficulty parameter of the PCM and the newly calculated LPCM were correlated and the minimal correlation, the median, the mean, the 95th percentile and the maximum correlation determined for each imputed dataset as can be seen in Table 54 to Table 56. Missing values could not be calculated due to the properties of the artificially generated design matrix.

Table 54

Descriptive statistics for the correlations obtained from simulated weight matrices for 20% occupancy – Item Set 4 - 6

$\%_1$	Min	Median	Mean	95%	Max
20	-0.3301	0.0064	0.0162	0.2372	0.3551
20	-0.3760	0.0048	-0.0016	0.1959	0.3971
20	-0.3505	0.0277	0.0260	0.2419	0.3832

Note. $\%_1$ = occupancy with 1's in the weight matrix; Min = minimum correlation, 95% = 95th percentile; Max = maximum correlation.

Table 55

Descriptive statistics for the correlations obtained from simulated weight matrices for 45% occupancy – Item Set 4 - 6

$\%_1$	Min	Median	Mean	95%	Max
45	-0.2566	0.0208	0.0278	0.2574	0.3238

Note. $\%_1$ = occupancy with 1's in the weight matrix; Min = minimum correlation, 95% = 95th percentile; Max = maximum correlation.

Table 56

Descriptive statistics for the correlations obtained from simulated weight matrices for 70% occupancy – Item Set 4 - 6

$\%_1$	Min	Median	Mean	95%	Max
70	-0.2662	0.0956	0.0956	0.3245	0.4024
70	-0.2940	0.1100	0.1033	0.3092	0.4132

Note. $\%_1$ = occupancy with 1's in the weight matrix; Min = minimum correlation, 95% = 95th percentile; Max = maximum correlation.

In a second simulation, q-matrices were permutated. The results of the descriptive statistics for each imputed dataset can be seen in Table 57.

Table 57

Descriptive statistics for the correlations obtained from permutated simulated weight matrices – Item Set 4 - 6

Min	Median	Mean	95%	Max
.4665	.7857	.7537	.8657	.8680
.5096	.8267	.7845	.8967	.8973
.5241	.8414	.8004	.9046	.9057
.4879	.8114	.7823	.8799	.8807
.5203	.8379	.8024	.9045	.9054

Note. Min = minimum correlation, 95% = 95th percentile; Max = maximum correlation.

Table 58

Item difficulty parameter for the PCM and the LPCM - Item Set 4 - 6

Item	Category	Item difficulty parameter PCM	Item difficulty parameter LPCM
F101	1	-2.429	-2.868
F201	1	-2.988	-2.868
F301	1	-2.843	-2.868
F101	1	-0.524	-2.868
F102	2	-1.684	-1.560
F201	1	0.210	-2.868
F202	2	-1.400	-1.560
F301	1	-0.707	-2.868
F302	2	-0.315	-1.560
F101	1	-1.239	-2.868
F102	2	-1.819	-1.560
F103	3	-1.521	0.487
F201	1	-1.435	-2.868
F202	2	-0.792	-1.560

continued

continued

Item	Category	Item difficulty parameter PCM	Item difficulty parameter LPCM
F203	3	0.946	0.487
F301	1	-1.603	-2.868
F302	2	-1.410	-1.560
F303	3	-1.615	0.487
F101	1	-0.874	-2.868
F102	2	-0.037	-1.560
F103	3	0.259	0.487
F104	4	0.809	2.732
F201	1	-0.815	-2.868
F202	2	-0.624	-1.560
F203	3	1.058	0.487
F204	4	1.799	2.732
F301	1	-0.391	-2.868
F302	2	-0.416	-1.560
F303	3	-0.008	0.487
F304	4	1.671	2.732
F101	1	-0.426	-2.868
F102	2	-0.018	-1.560
F103	3	0.977	0.487
F104	4	1.648	2.732
F105	5	3.346	5.550
F201	1	-1.272	-2.868
F202	2	-0.781	-1.560
F203	3	0.604	0.487
F204	4	1.272	2.732
F205	5	3.600	5.550
F301	1	-1.090	-2.868
F302	2	-0.990	-1.560

continued

continued

Item	Category	Item difficulty parameter PCM	Item difficulty parameter LPCM
F303	3	0.492	0.487
F304	4	1.133	2.732
F305	5	2.994	5.550
F101	1	-1.009	-2.868
F102	2	-1.153	-1.560
F103	3	-0.456	0.487
F104	4	0.096	2.732
F105	5	2.186	5.550
F106	6	3.391	7.810
F301	1	-0.665	-2.868
F302	2	-0.314	-1.560
F303	3	0.126	0.487
F304	4	1.521	2.732
F305	5	2.039	5.550
F306	6	3.487	7.810

Working memory verbal.

Item Set 7. There was no option to calculate the item parameters due to the response pattern.

Item Set 8. Due to missing response patterns, no LRT could be computed. Martin-Löf-test showed a not significant result ($p = .29$) as well as T_{11}-statistic ($p = .27$). IFA was not significant ($p = .23$). LLTM and RM item difficulty parameter correlated highly with $r = .98$ ($p < .001$).

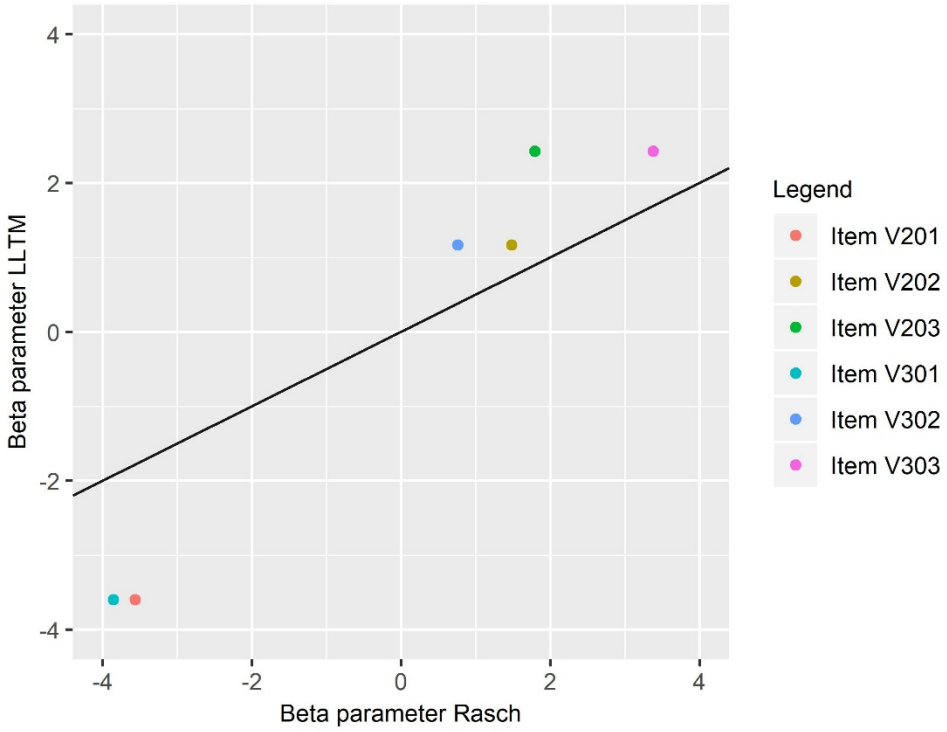


Figure 27. RM and LLTM beta parameter of Item Set 8.

Table 59

Item difficulty parameter for the RM and the LLTM - Item Set 8

Item	Item difficulty parameter Rasch	Item difficulty parameter LLTM
V201	-3.563	-3.597
V301	-3.857	-3.597
V202	1.484	1.171
V302	0.763	1.171
V203	1.792	2.427
V303	3.381	2.427

For the PCM, LRT showed no significance ($p = .67$)[22]. Martin-Löf-test could not be computed. IFA was not significant as well ($p = .19$). Item difficulty parameters of PCM and LPCM correlated with $r = .73$ ($p < .01$). Item difficulty parameters for the PCM and the LPCM without an a priori defined q-matrix correlated with $r = .92$ ($p < .001$).

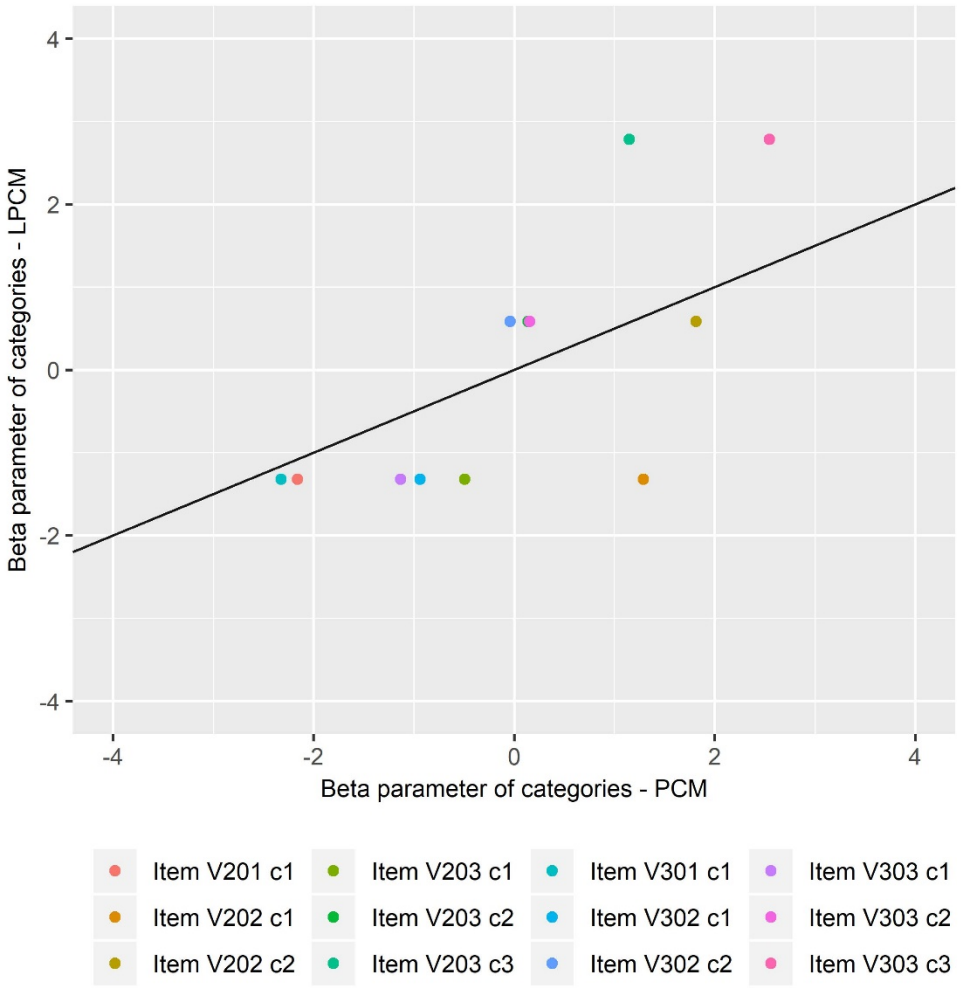


Figure 28. PCM and LPCM beta parameter of categories of Item Set 8.

In a simulation, random q-matrices with different ratios of 0's and 1's were generated and the LPCM calculated with those. The item difficulty parameter of the PCM

[22] Items F202, F302, F203 and F303 had to be removed for the test because of missing response patterns within subgroups with split criterion mean.

and the newly calculated LPCM were correlated and the minimal correlation, the median, the mean, the 95th percentile and the maximum correlation determined as can be seen in Table 60. Missing values could not be calculated due to the properties of the artificially generated design matrix.

Table 60

Descriptive statistics for the correlations obtained from simulated weight matrices – Item Set 8

$\%_1$	Min	Median	Mean	95%	Max
25	-.4634	.1997	.1902	.6130	.7293
30	-.3991	.2701	.2547	.5722	.7012
35	-.3298	.2680	.2448	.6882	.8387
40	-.2578	.2666	.2508	.6508	.8938
45	-.5377	.2470	.2244	.6326	.8278
50	-.4918	.2189	.2470	.6535	.8559
55	-.4032	.2903	.3016	.7615	.8273
60	-.3488	.3200	.2835	.5860	.7940

Note. $\%_1$ = occupancy with 1's in the weight matrix; Min = minimum correlation, 95% = 95th percentile; Max = maximum correlation.

In a second simulation, q-matrices were permutated. The results of the descriptive statistics can be seen in Table 61.

Table 61

Descriptive statistics for the correlations obtained from permutated simulated weight matrices – Item Set 8

Min	Median	Mean	95%	Max
.1354	.5671	.5104	.6878	.7264

Note. Min = minimum correlation, 95% = 95th percentile; Max = maximum correlation.

Table 62

Item difficulty parameter for the PCM and the LPCM - Item Set 8

Item	Category	Item difficulty parameter PCM	Item difficulty parameter LPCM
V201	1	-2.162	-1.320
V301	1	-2.323	-1.320
V202	1	1.290	-1.320
V202	2	1.813	0.586
V302	1	-0.937	-1.320
V302	2	-0.040	0.586
V203	1	-0.492	-1.320
V203	2	0.138	0.586
V203	3	1.146	2.786
V303	1	-1.133	-1.320
V303	2	0.155	0.586
V303	3	2.544	2.786

Item Set 9. LRT was not significant ($p = .63$)[23] as was Martin-Löf-test ($p = .10$). The χ^2/df of the LRT was 0.46. IFA was not significant ($p = .49$). Both item difficulty parameters correlated highly ($r = .99$; $p < .001$).

[23] Items V102, V303, V103, V104 and V305 had to be removed for the test because of missing response patterns within subgroups with split criterion mean.

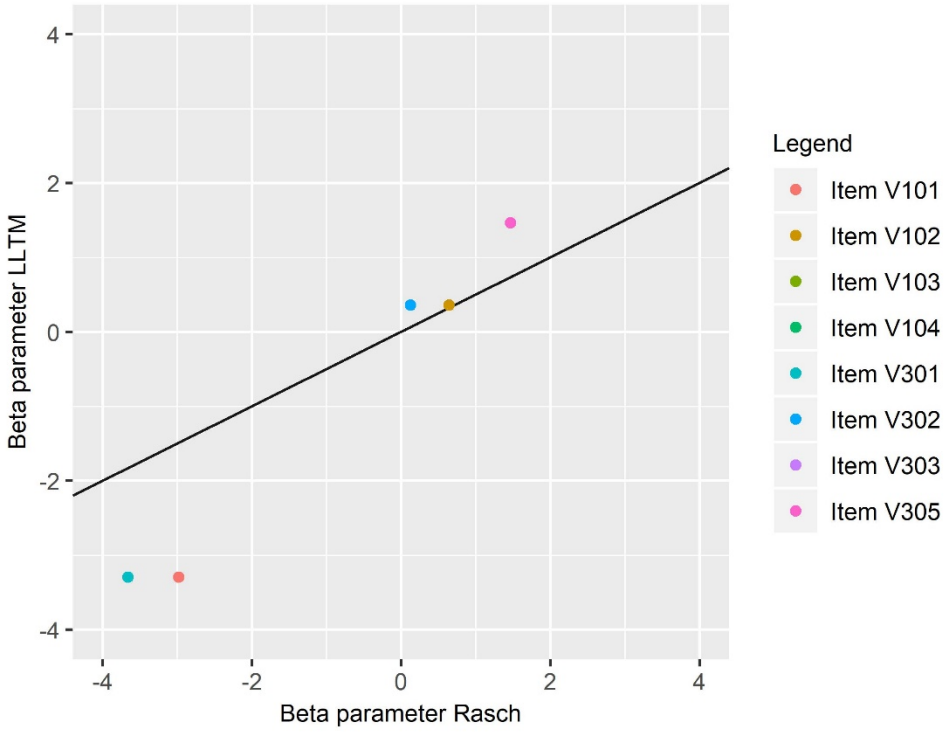


Figure 29. RM and LLTM beta parameter of Item Set 9.[24]

Table 63

Item difficulty parameter for the RM and the LLTM - Item Set 9

Item	Item difficulty parameter Rasch	Item difficulty parameter LLTM
V301	-3.660	-3.294
V101	-2.981	-3.294
V302	0.126	0.361
V102	0.641	0.361
V303	1.468	1.467
V103	1.468	1.467
V104	1.468	1.467
V305[25]	1.468	1.467

[24] Items not appearing in the plot share the identical beta parameter for LLTM and RM until the third decimal.

[25] Item V305 is included in the LLTM, but not in the LPCM due to incomplete response patterns.

For the PCM, LRT showed no significance ($p = .74$)[26] as well as Martin-Löf-test ($p = .99$). IFA was significant ($p < .05$). Item difficulty parameters of PCM and LPCM correlated with $r = .72$ ($p < .01$). Item difficulty parameters for the PCM and the LPCM without an a priori defined q-matrix correlated with $r = .86$ ($p < .001$).

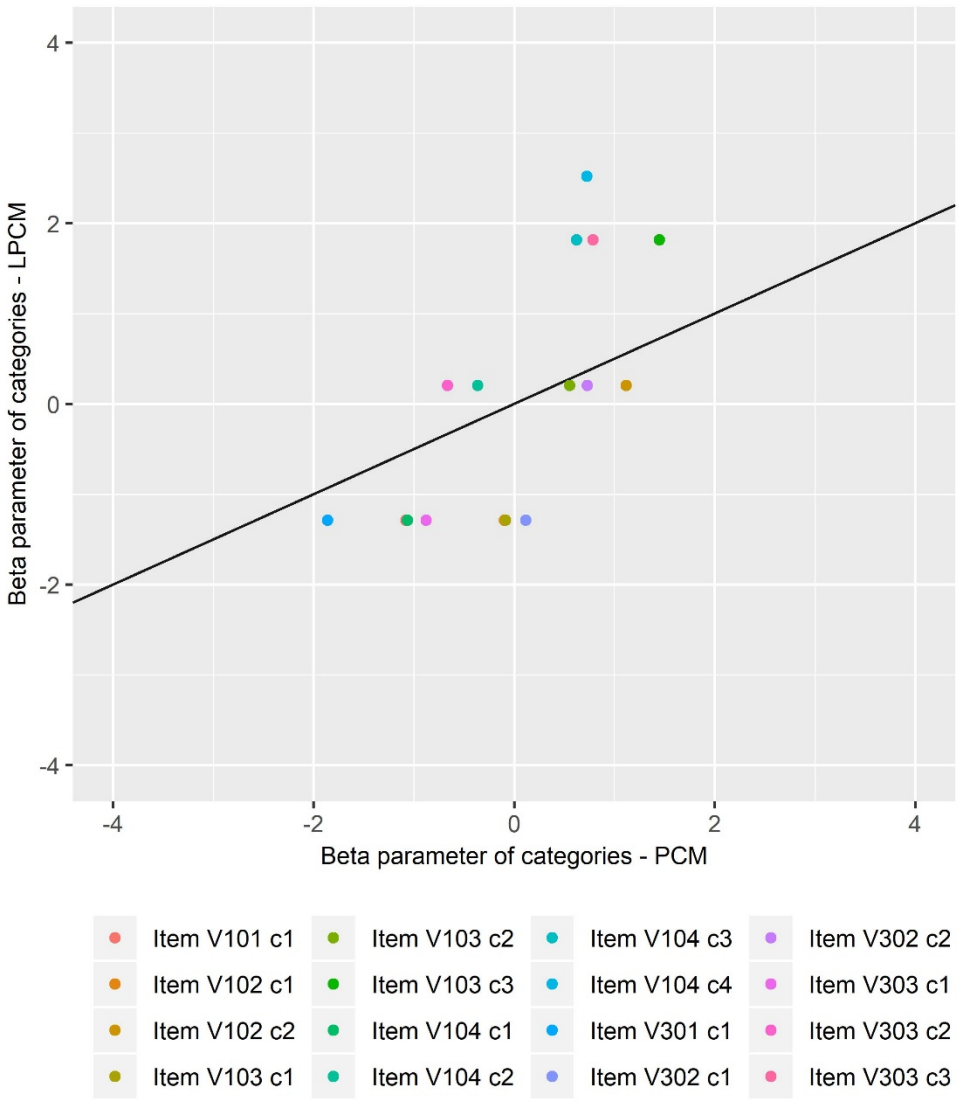


Figure 30. PCM and LPCM beta parameter of categories of Item Set 9.

[26] Items V302, V102, V303, V103 and V104 had to be removed for the test because of missing response patterns within subgroups with split criterion mean.

In a simulation, random q-matrices with different ratios of 0's and 1's were generated and the LPCM calculated with those. The item difficulty parameter of the PCM and the newly calculated LPCM were correlated and the minimal correlation, the median, the mean, the 95th percentile and the maximum correlation determined as can be seen in Table 64. Missing values could not be calculated due to the properties of the artificially generated design matrix.

Table 64

Descriptive statistics for the correlations obtained from simulated weight matrices – Item Set 9

$\%_1$	Min	Median	Mean	95%	Max
20	-.5764	.2301	.2204	.5737	.7424
25	-.3014	.2352	.2347	.6077	.6898
30	-.5062	.2419	.2384	.6122	.7479
35	-.2820	.2404	.2098	.6064	.7673
40	-.4080	.2781	.2488	.6290	.7341
45	-.4947	.2884	.2417	.6124	.8370
50	-.3763	.2497	.2372	.6278	.7820
55	-.1882	.3105	.2803	.6365	.7666
60	-.2489	.3284	.3114	.6322	.8348

Note. $\%_1$ = occupancy with 1's in the weight matrix; Min = minimum correlation, 95% = 95th percentile; Max = maximum correlation.

In a second simulation, q-matrices were permutated. The results of the descriptive statistics can be seen in Table 65.

Table 65

Descriptive statistics for the correlations obtained from permutated simulated weight matrices – Item Set 9

Min	Median	Mean	95%	Max
.2447	.5723	.4934	.6715	.7201

Note. Min = minimum correlation, 95% = 95th percentile; Max = maximum correlation.

Table 66

Item difficulty parameter for the PCM and the LPCM - Item Set 9

Item	Category	Item difficulty parameter PCM	Item difficulty parameter LPCM
V301	1	-1.861	-1.286
V101	1	-1.081	-1.286
V302	1	0.114	-1.286
V302	2	0.730	0.206
V102	1	-0.099	-1.286
V102	2	1.119	0.206
V303	1	-0.879	-1.286
V303	2	-0.664	0.206
V303	3	0.785	1.817
V103	1	-0.086	-1.286
V103	2	0.552	0.206
V103	3	1.450	1.817
V104	1	-1.064	-1.286
V104	2	-0.364	0.206
V104	3	0.623	1.817
V104	4	0.726	2.521

Item Set 7 - 9. For Item Set 7 to 9, LRT was not significant ($p = .74$). However, Martin-Löf-test could not be computed. IFA was significant ($p < .05$). Item difficulty parameters of the RM and the LLTM correlated with $r = .99$ ($p < .001$).

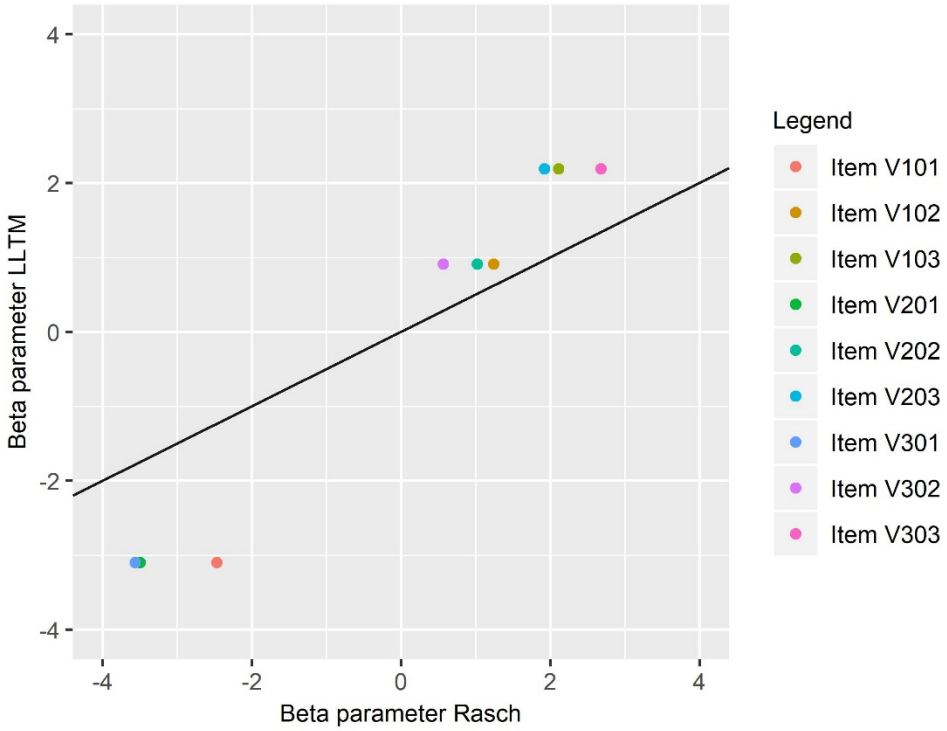


Figure 31. RM and LLTM beta parameter of Item Set 7 – 9.

Table 67

Item difficulty parameter for the RM and the LLTM - Item Set 7-9

Item	Item difficulty parameter Rasch	Item difficulty parameter LLTM
V101	-2.470	-3.099
V201	-3.499	-3.099
V301	-3.567	-3.099
V102	1.239	0.910
V202	1.019	0.910
V302	0.567	0.910
V103	2.109	2.190
V203	1.921	2.190
V303	2.681	2.190
V104	-2.470	-3.099
V305[27]	-3.499	-3.099

For the PCM, LRT could not be calculated. However, Martin-Löf-test showed no significance ($p = .16$). Item difficulty parameters of PCM and LPCM correlated with $r = .79$ ($p < .001$). Item difficulty parameters for the PCM and the LPCM without an a priori defined q-matrix correlated with $r = 1$ ($p < .001$).

[27] Item V305 is included in the LLTM, but not in the LPCM due to incomplete response patterns.

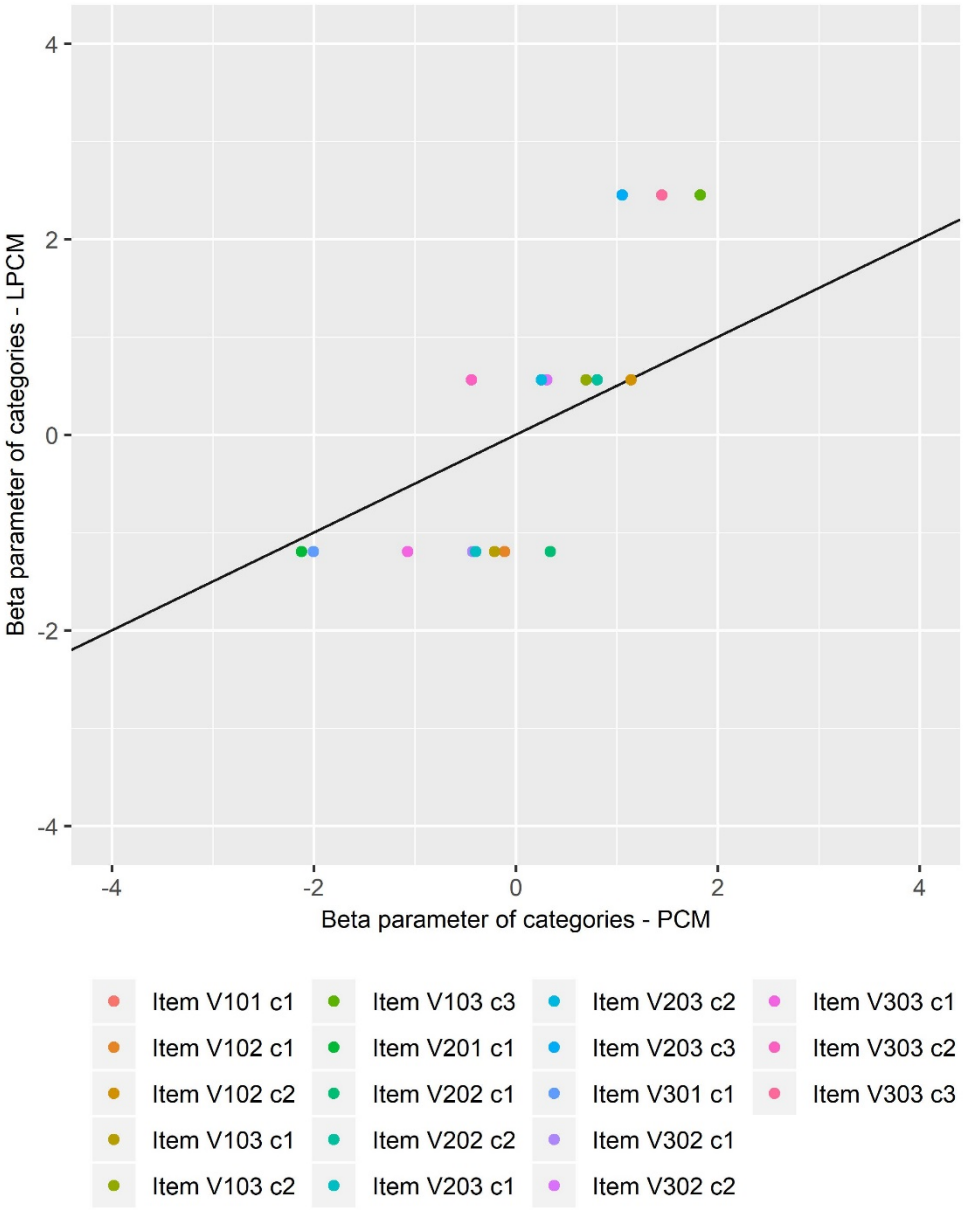


Figure 32. PCM and LPCM beta parameter of categories of Item Set 7 – 9.

In a simulation, random q-matrices with different ratios of 0's and 1's were generated and the LPCM calculated with those. The item difficulty parameter of the PCM and the newly calculated LPCM were correlated and the minimal correlation, the median, the mean, the 95th percentile and the maximum correlation determined for each imputed dataset as can be seen in Table 68 to Table 78. Missing values could not be calculated due to the properties of the artificially generated design matrix.

Table 68

Descriptive statistics for the correlations obtained from simulated weight matrices for 20% occupancy – Item Set 7 - 9

$\%_1$	Min	Median	Mean	95%	Max
20	-.4643	.1563	.1252	.4687	.6251
20	-.6608	.1395	.1113	.3968	.6204
20	-.3908	.1205	.1143	.4695	.5828
20	-.4657	.1231	.1113	.4091	.5955
20	-.4295	.1004	.1226	.5210	.7458

Note. $\%_1$ = occupancy with 1's in the weight matrix; Min = minimum correlation, 95% = 95th percentile; Max = maximum correlation.

Table 69

Descriptive statistics for the correlations obtained from simulated weight matrices for 25% occupancy – Item Set 7 - 9

$\%_1$	Min	Median	Mean	95%	Max
25	-.4427	.1657	0.1386	.4877	.6242
25	-.4631	.1894	.1329	.4638	.6753
25	-.3953	.1605	.1554	.4956	.5662
25	-.4745	.1432	.1357	.5381	.6285
25	-.5731	.0663	.0846	.4305	.7001

Note. $\%_1$ = occupancy with 1's in the weight matrix; Min = minimum correlation, 95% = 95th percentile; Max = maximum correlation.

Table 70

Descriptive statistics for the correlations obtained from simulated weight matrices for 30% occupancy – Item Set 7 - 9

$\%_1$	Min	Median	Mean	95%	Max
30	-.3647	.1986	.1607	.4530	.5345
30	-.2840	.1503	.1529	.4620	.5907
30	-.2882	.1625	.1323	.4183	.5133
30	-.3103	.1158	.1329	.4695	.6726
30	-.3592	.1504	.1326	.4655	.7295

Note. $\%_1$ = occupancy with 1's in the weight matrix; Min = minimum correlation, 95% = 95th percentile; Max = maximum correlation.

Table 71

Descriptive statistics for the correlations obtained from simulated weight matrices for 35% occupancy – Item Set 7 - 9

$\%_1$	Min	Median	Mean	95%	Max
35	-.4086	.1853	.1456	.4653	.5908
35	-.3608	.1164	.1299	.5244	.6689
35	-.3319	.2101	.1882	.5015	.6130
35	-.4999	.1539	.1351	.5083	.6634
35	-.5903	.1486	.1361	.4968	.6160

Note. $\%_1$ = occupancy with 1's in the weight matrix; Min = minimum correlation, 95% = 95th percentile; Max = maximum correlation.

Table 72

Descriptive statistics for the correlations obtained from simulated weight matrices for 40% occupancy – Item Set 7 - 9

$\%_1$	Min	Median	Mean	95%	Max
40	-.4537	.1972	.1561	.4940	.6610
40	-.2957	.1175	.1455	.5109	.6097
40	-.5016	.1252	.1376	.5064	.6617
40	-.3445	.1905	.1770	.4989	.6206
40	-.3221	.1813	.1736	.5218	.6200

Note. $\%_1$ = occupancy with 1's in the weight matrix; Min = minimum correlation, 95% = 95th percentile; Max = maximum correlation.

Table 73

Descriptive statistics for the correlations obtained from simulated weight matrices for 45% occupancy – Item Set 7 - 9

$\%_1$	Min	Median	Mean	95%	Max
45	-.3472	.1929	.1746	.5489	.7802
45	-.3003	.1746	.1535	.5736	.8543
45	-.3362	.1628	.1588	.5426	.5897
45	-.2355	.2137	.1890	.4737	.6546
45	-.3380	.1487	.1562	.5040	.6815

Note. $\%_1$ = occupancy with 1's in the weight matrix; Min = minimum correlation, 95% = 95th percentile; Max = maximum correlation.

Table 74

Descriptive statistics for the correlations obtained from simulated weight matrices for 50% occupancy – Item Set 7 - 9

$\%_1$	Min	Median	Mean	95%	Max
50	-.3481	.2153	.2007	.5367	.6791
50	-.2925	.1957	.2156	.5469	.6846
50	-.2568	.1793	.1699	.5121	.6664
50	-.3351	.1707	.1823	.5530	.6656
50	-.4022	.1393	.1606	.5442	.6914

Note. $\%_1$ = occupancy with 1's in the weight matrix; Min = minimum correlation, 95% = 95th percentile; Max = maximum correlation.

Table 75

Descriptive statistics for the correlations obtained from simulated weight matrices for 55% occupancy – Item Set 7 - 9

$\%_1$	Min	Median	Mean	95%	Max
55	-.2452	.2087	.2113	.5167	.6305
55	-.4288	.1982	.1827	.5368	.6298
55	-.3621	.1575	.1815	.5948	.6350
55	-.3331	.1810	.1769	.4911	.6984
55	-.2401	.1560	.1684	.5228	.5823

Note. $\%_1$ = occupancy with 1's in the weight matrix; Min = minimum correlation, 95% = 95th percentile; Max = maximum correlation.

Table 76

Descriptive statistics for the correlations obtained from simulated weight matrices for 60% occupancy – Item Set 7 - 9

$\%_1$	Min	Median	Mean	95%	Max
60	-.2742	.2356	.2199	.6101	.7174
60	-.2865	.2372	.2166	.5251	.6583
60	-.3021	.2371	.2175	.5928	.7774
60	-.2395	.2076	.2157	.5413	.6278
60	-.2873	.2076	.2246	.5633	.6426

Note. $\%_1$ = occupancy with 1's in the weight matrix; Min = minimum correlation, 95% = 95th percentile; Max = maximum correlation.

Table 77

Descriptive statistics for the correlations obtained from simulated weight matrices for 65% occupancy – Item Set 7 - 9

$\%_1$	Min	Median	Mean	95%	Max
65	-.1989	.2590	.2426	.5481	.7114
65	-.3225	.2174	.2061	.6012	.6238
65	-.3526	.2065	.1966	.4937	.6769
65	-.2305	.2197	.2096	.5053	.5625
65	-.2521	.2002	.2089	.5506	.5954

Note. $\%_1$ = occupancy with 1's in the weight matrix; Min = minimum correlation, 95% = 95th percentile; Max = maximum correlation.

Table 78

Descriptive statistics for the correlations obtained from simulated weight matrices for 70% occupancy – Item Set 7 - 9

$\%_1$	Min	Median	Mean	95%	Max
70	-.2486	.2156	.2175	.4972	.6346
70	-.2618	.2173	.2234	.5635	.7466
70	-.3206	.2484	.2297	.5487	.7772
70	-.1345	.2718	.2570	.5291	.6302

Note. $\%_1$ = occupancy with 1's in the weight matrix; Min = minimum correlation, 95% = 95th percentile; Max = maximum correlation.

In a second simulation, q-matrices were permutated. The results of the descriptive statistics for each imputed dataset can be seen in Table 79.

Table 79

Descriptive statistics for the correlations obtained from permutated simulated weight matrices – Item Set 7 - 9

Min	Median	Mean	95%	Max
.1242	.6327	.5824	.7825	.7825
.1920	.6170	.5921	.7976	.7976
.1834	.5628	.5395	.7654	.7654
.1278	.5653	.5439	.7665	.7665
.1244	.6172	.5627	.7882	.7882

Note. Min = minimum correlation, 95% = 95th percentile; Max = maximum correlation.

Table 80

Item difficulty parameter for the PCM and the LPCM - Item Set 7 - 9

Item	Category	Item difficulty parameter PCM	Item difficulty parameter LPCM
V101	1	-1.072	-1.192
V201	1	-2.121	-1.192
V301	1	-2.006	-1.192
V101	1	-0.111	-1.192
V102	2	1.139	0.561
V201	1	0.342	-1.192
V202	2	0.806	0.561
V301	1	-0.430	-1.192
V302	2	0.306	0.561
V101	1	-0.212	-1.192
V102	2	0.696	0.561
V103	3	1.824	2.455
V201	1	-0.399	-1.192
V202	2	0.254	0.561
V203	3	1.052	2.455
V301	1	-1.073	-1.192
V302	2	-0.439	0.561
V303	3	1.445	2.455

Item Set 10. LRT was not significant (p = .34)[28] as well as Martin-Löf-test (p = .72) as the T_{11}-statistic (p = .39). The χ^2/df of the LRT was 1.12. IFA was not significant (p = .48). The item difficulty parameter of the LLTM and the RM correlated highly with r = .94 (p < .001).

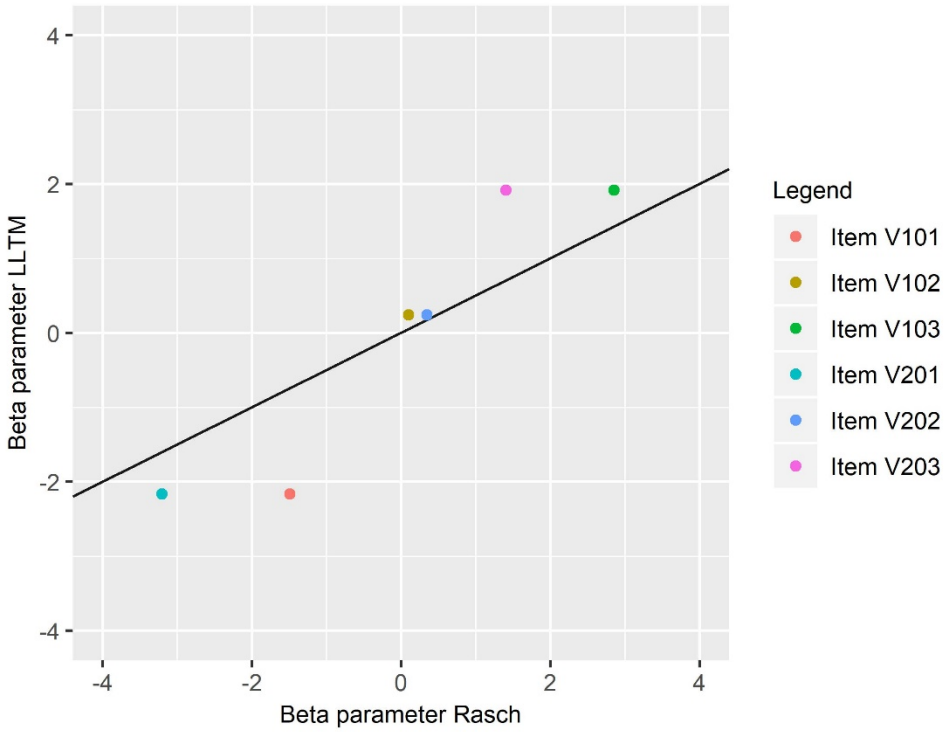


Figure 33. RM and LLTM beta parameter of Item Set 10.

[28] Items V103 and V203 had to be removed for the test because of missing response patterns within subgroups with split criterion mean.

Table 81

Item difficulty parameter for the RM and the LLTM - Item Set 10

Item	Item difficulty parameter Rasch	Item difficulty parameter LLTM
V101	-1.492	-2.162
V201	-3.209	-2.162
V102	0.098	0.241
V202	0.343	0.241
V103	2.852	1.921
V203	1.407	1.921

For the PCM, LRT showed no significance ($p = .25$)[29] as well as Martin-Löf-test ($p = .46$). IFA was not significant ($p = .91$). Item difficulty parameters of PCM and LPCM correlated with $r = .82$ ($p < .001$). Item difficulty parameters for the PCM and the LPCM without an a priori defined q-matrix correlated with $r = .98$ ($p < .001$).

[29] Items V202, V103 and V203 had to be removed for the test because of missing response patterns within subgroups with split criterion mean.

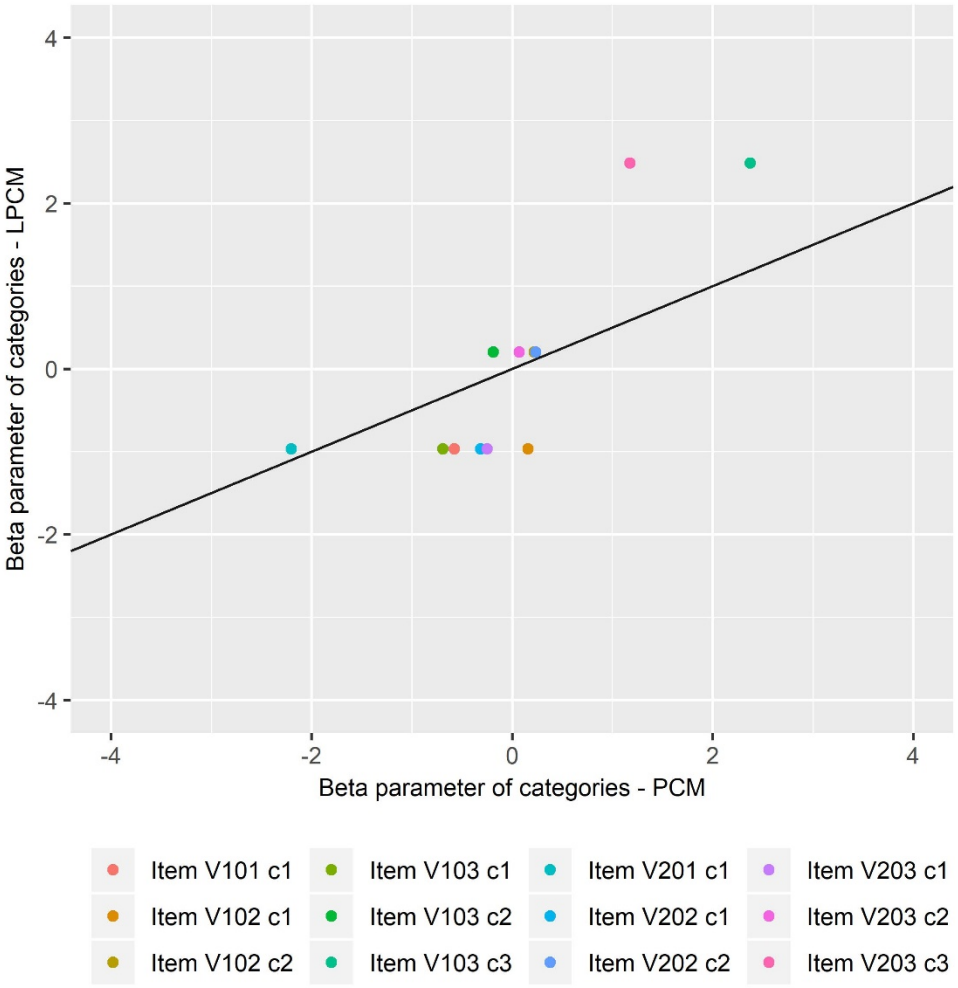


Figure 34. PCM and LPCM beta parameter of categories of Item Set 10.

In a simulation, random q-matrices with different ratios of 0's and 1's were generated and the LPCM calculated with those. The item difficulty parameter of the PCM and the newly calculated LPCM were correlated and the minimal correlation, the median, the mean, the 95th percentile and the maximum correlation determined as can be seen in Table 82. Missing values could not be calculated due to the properties of the artificially generated design matrix.

Table 82

Descriptive statistics for the correlations obtained from simulated weight matrices – Item Set 10

$\%_1$	Min	Median	Mean	95%	Max
20	-.4439	.1433	.1964	.6787	.7143
25	-.5319	.1917	.2293	.6978	.7686
30	-.3562	.1223	.1554	.5882	.7387
35	-.3655	.1931	.1885	.6353	.7252
40	-.2639	.1958	.2026	.5750	.6332
45	-.2682	.3035	.2787	.6094	.7466
50	-.5162	.2758	.2533	.6203	.6765
55	-.2040	.3336	.3039	.7022	.7614
60	-.4628	.3053	.2900	.6024	.7637
65	-.2000	.3021	.3193	.6740	.8012

Note. $\%_1$ = occupancy with 1's in the weight matrix; Min = minimum correlation, 95% = 95th percentile; Max = maximum correlation.

In a second simulation, q-matrices were permutated. The results of the descriptive statistics can be seen in Table 83.

Table 83

Descriptive statistics for the correlations obtained from permutated simulated weight matrices – Item Set 10

Min	Median	Mean	95%	Max
.1936	.7388	.5947	.8204	.8225

Note. Min = minimum correlation, 95% = 95th percentile; Max = maximum correlation.

Table 84

Item difficulty parameter for the PCM and the LPCM - Item Set 10

Item	Category	Item difficulty parameter PCM	Item difficulty parameter LPCM
V101	1	-0.577	-0.966
V201	1	-2.205	-0.966
V102	1	0.158	-0.966
V102	2	0.218	0.205
V202	1	-0.316	-0.966
V202	2	0.234	0.205
V103	1	-0.691	-0.966
V103	2	-0.190	0.205
V103	3	2.373	2.490
V203	1	-0.249	-0.966
V203	2	0.071	0.205
V203	3	1.173	2.490

Item Set 11. Due to insufficient response patterns, LRT could not be calculated. Martin-Löf-test (p = .85) and T_{11}-statistic was not significant (p = .38). IFA was not significant (p = .30). Item difficulty parameter of the LLTM and RM correlated (r = .99, p <.001).

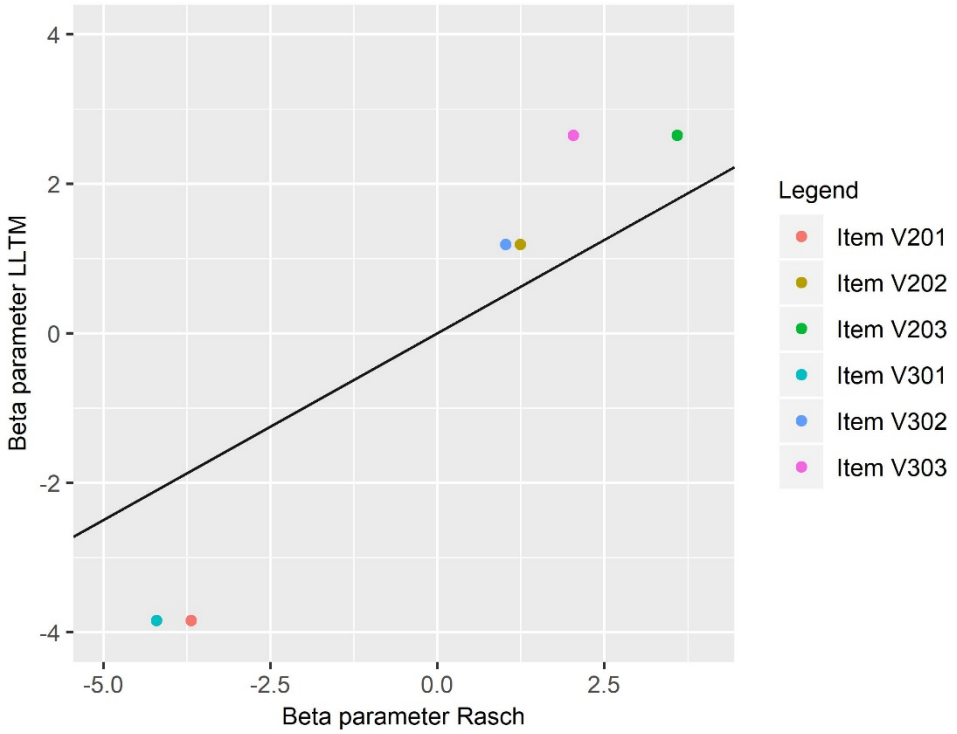


Figure 35. RM and LLTM beta parameter of Item Set 11.

Table 85

Item difficulty parameter for the RM and the LLTM - Item Set 11

Item	Item difficulty parameter Rasch	Item difficulty parameter LLTM
V201	-3.685	-3.842
V301	-4.203	-3.842
V202	1.241	1.192
V302	1.023	1.192
V203	3.589	2.650
V303	2.034	2.650

For the PCM, LRT showed no significance ($p = .85$)[30] as well as Martin-Löf-test ($p = .18$). IFA was significant ($p = .03$). Item difficulty parameters of PCM and LPCM correlated with $r = .88$ ($p < .001$). Item difficulty parameters for the PCM and the LPCM without an a priori defined q-matrix correlated with $r = .97$ ($p < .001$).

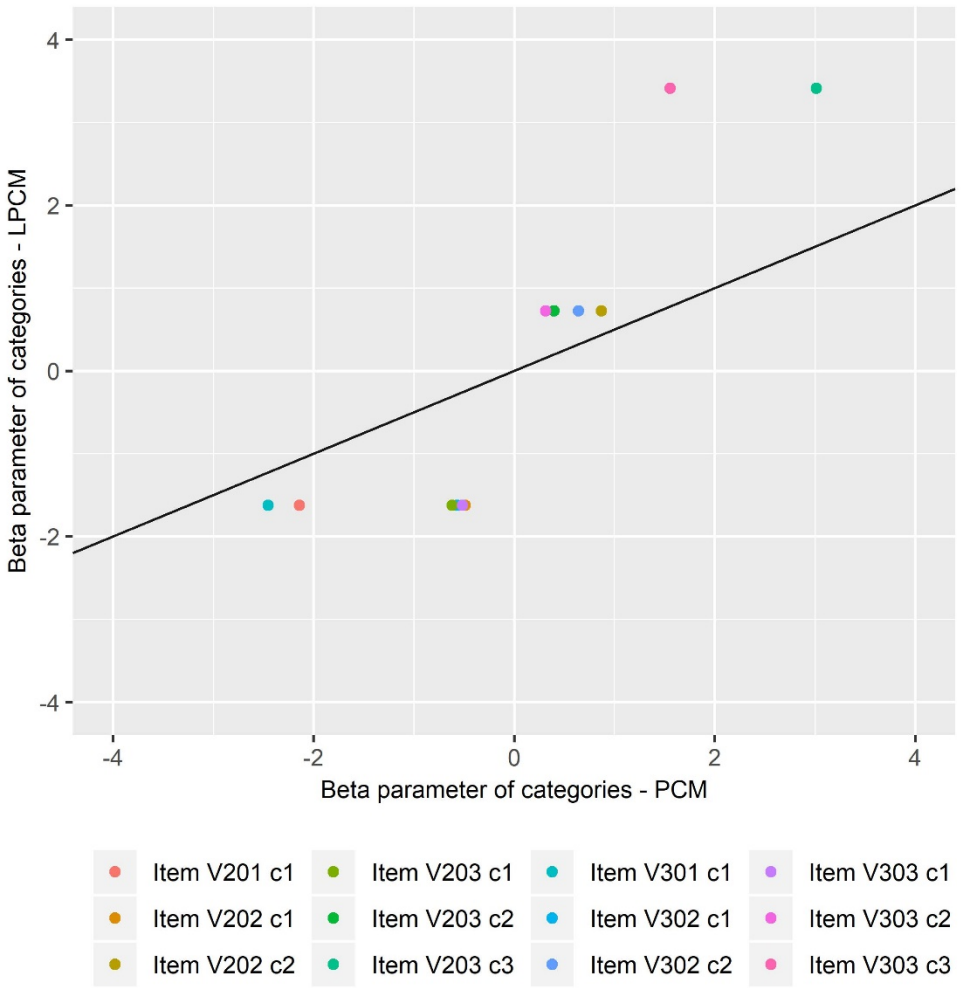


Figure 36. PCM and LPCM beta parameter of categories of Item Set 11.

In a simulation, random q-matrices with different ratios of 0's and 1's were generated and the LPCM calculated with those. The item difficulty parameter of the PCM

[30] Items V202, V302, V203 and V303 had to be removed for the test because of missing response patterns within subgroups with split criterion mean.

and the newly calculated LPCM were correlated and the minimal correlation, the median, the mean, the 95th percentile and the maximum correlation determined as can be seen in Table 86. Missing values could not be calculated due to the properties of the artificially generated design matrix.

Table 86

Descriptive statistics for the correlations obtained from simulated weight matrices – Item Set 11

$\%_1$	Min	Median	Mean	95%	Max
25	-.5751	.1360	.1810	.6839	.8456
30	-.4432	.1366	.1276	.6544	.8393
35	-.6248	.1564	.1663	.6819	.8225
40	-.6285	.1443	.1657	.6127	.8622
45	-.4871	.1410	.1479	.6209	.9206
50	-.4962	.1786	.1846	.6167	.7434
55	-.4353	.1858	.1978	.6218	.7597
60	-.3829	.2402	.2641	.7197	.8046
65	-.3790	.2608	.2715	.6799	.8175

Note. $\%_1$ = occupancy with 1's in the weight matrix; Min = minimum correlation, 95% = 95th percentile; Max = maximum correlation.

In a second simulation, q-matrices were permutated. The results of the descriptive statistics can be seen in Table 87.

Table 87

Descriptive statistics for the correlations obtained from permutated simulated weight matrices – Item Set 11

Min	Median	Mean	95%	Max
-.0076	.6931	.5384	.7859	.8838

Note. Min = minimum correlation, 95% = 95th percentile; Max = maximum correlation.

Table 88

Item difficulty parameter for the PCM and the LPCM - Item Set 11

Item	Category	Item difficulty parameter PCM	Item difficulty parameter LPCM
V201	1	-2.141	-1.621
V301	1	-2.454	-1.621
V202	1	-0.488	-1.621
V202	2	0.871	0.725
V302	1	-0.568	-1.621
V302	2	0.640	0.725
V203	1	-0.620	-1.621
V203	2	0.398	0.725
V203	3	3.011	3.414
V303	1	-0.518	-1.621
V303	2	0.315	0.725
V303	3	1.553	3.414

Item Set 12. LRT was not significant (p = .82)[31] as well as Martin-Löf-test (p = .93) and the T_{11}-statistic (p = .10). The χ^2/df of the LRT was 0.38. IFA was significant (p = .02). Parameter estimations of the item difficulty of the LLTM and the RM correlated with r = .97 (p < .001). Item difficulty parameters for the PCM and the LPCM without an a priori defined q-matrix correlated with r = .96 (p < .001).

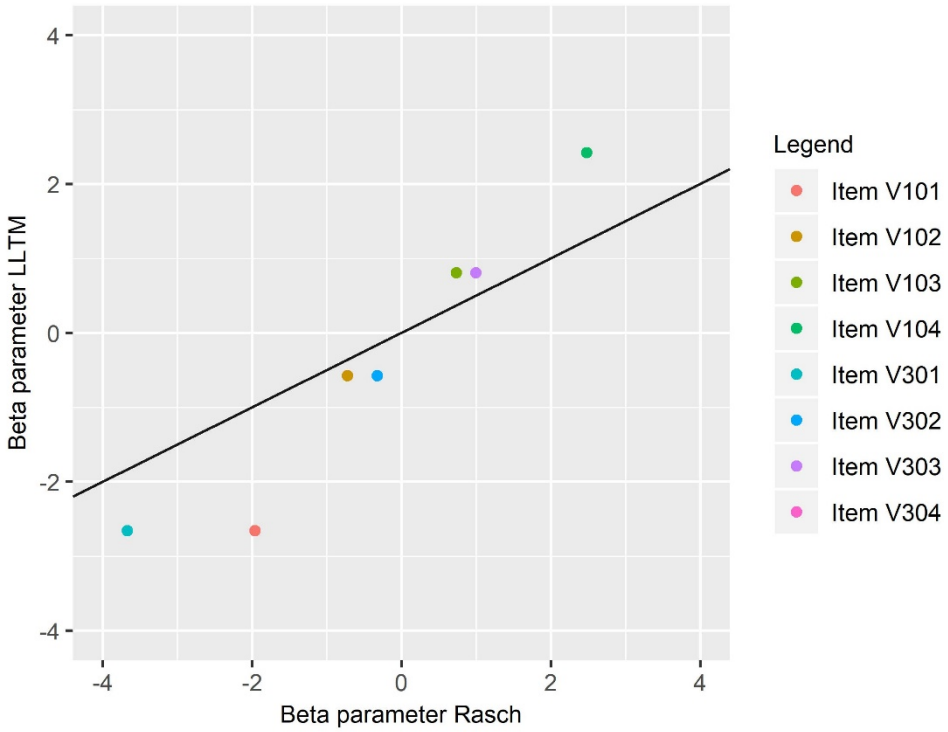


Figure 37. RM and LLTM beta parameter of Item Set 12.

[31] Items V103, V304 and V104 had to be removed for the test because of missing response patterns within subgroups with split criterion median.

Table 89

Item difficulty parameter for the RM and the LLTM - Item Set 12

Item	Item difficulty parameter Rasch	Item difficulty parameter LLTM
V301	-3.674	-2.658
V101	-1.960	-2.658
V302	-0.326	-0.575
V102	-0.725	-0.575
V303	0.994	0.807
V103	0.734	0.807
V304	2.479	2.426
V104	2.479	2.426

For the PCM, LRT showed no significance ($p = .96$)[32] as well as Martin-Löf-test ($p = 1.0$). IFA was not significant ($p = .40$). Item difficulty parameters of PCM and LPCM correlated with $r = .91$ ($p < .001$).

[32] Items V303, V103, V304 and V104 had to be removed for the test because of missing response patterns within subgroups with split criterion mean.

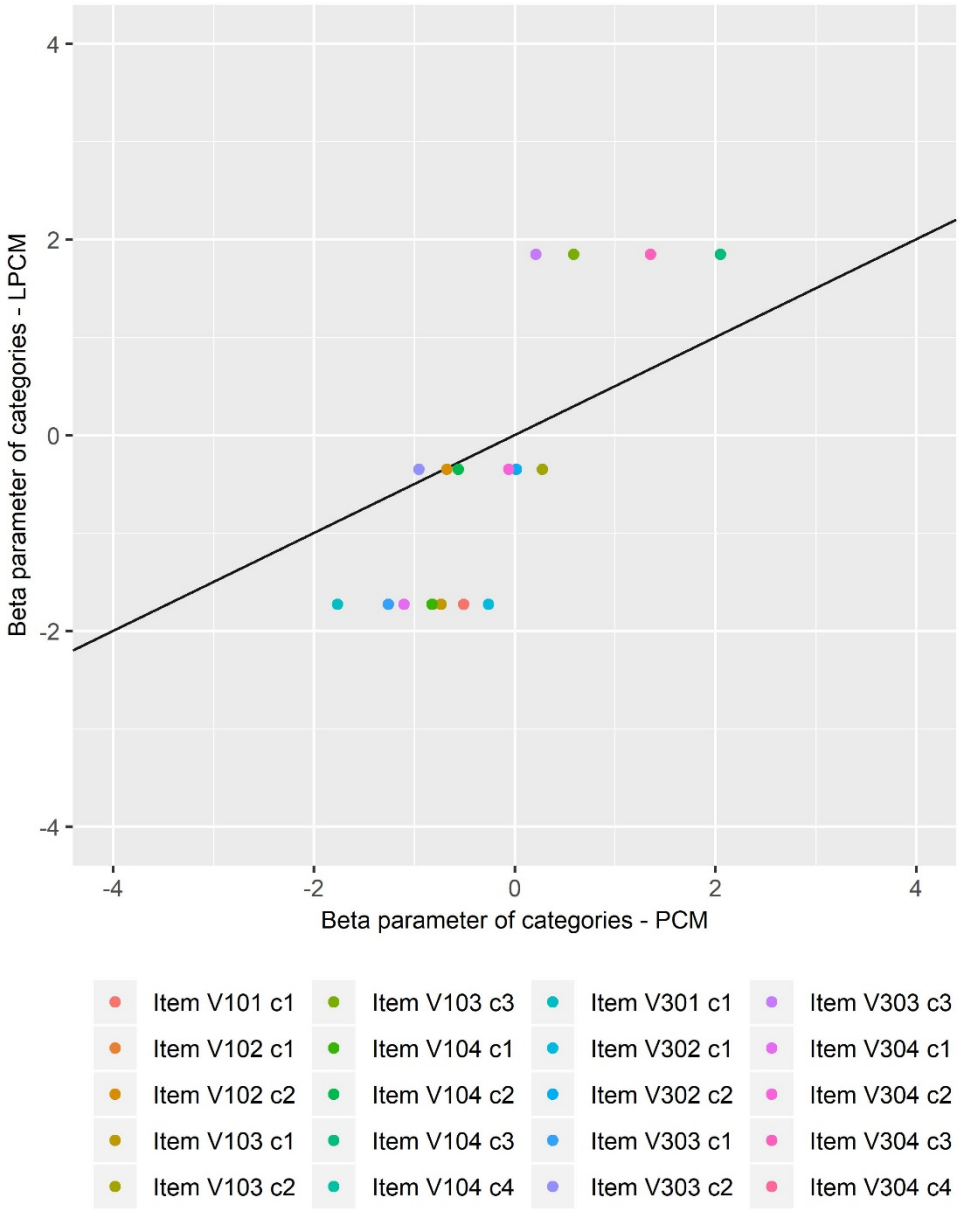


Figure 38. PCM and LPCM beta parameter of categories of Item Set 12.

In a simulation, random q-matrices with different ratios of 0's and 1's were generated and the LPCM calculated with those. The item difficulty parameter of the PCM and the newly calculated LPCM were correlated and the minimal correlation, the median, the mean, the 95th percentile and the maximum correlation determined as can be seen in Table 90.

Table 90

Descriptive statistics for the correlations obtained from simulated weight matrices – Item Set 12

$\%_1$	Min	Median	Mean	95%	Max
20	-.4732	.1345	.1431	.6211	.7087
25	-.4194	.1603	.1462	.4842	.6922
30	-.4227	.0449	.0499	.4467	.5224
35	-.4265	.1221	.1067	.5165	.6593
40	-.3523	.1301	.1266	.5600	.7276
45	-.4288	.1390	.1274	.5189	.6645
50	-.5516	.1885	.1568	.5463	.6716
55	-.4264	.1865	.1558	.4644	.5409
60	-.2797	.1513	.1755	.5309	.7621
65	-.2915	.1812	.1781	.5473	.6024
70	-.3506	.1602	.1647	.4793	.7527

Note. $\%_1$ = occupancy with 1's in the weight matrix; Min = minimum correlation, 95% = 95th percentile; Max = maximum correlation.

In a second simulation, q-matrices were permutated. The results of the descriptive statistics can be seen in Table 91.

Table 91

Descriptive statistics for the correlations obtained from permutated simulated weight matrices – Item Set 12

Min	Median	Mean	95%	Max
.5901	.8111	.8017	.9132	.9199

Note. Min = minimum correlation, 95% = 95th percentile; Max = maximum correlation.

Table 92

Item difficulty parameter for the PCM and the LPCM - Item Set 12

Item	Category	Item difficulty parameter PCM	Item difficulty parameter LPCM
V301	1	-2.288	-1.639
V101	1	-2.981	-1.639
V302	1	0.013	-1.639
V302	2	-0.552	-0.138
V102	1	0.236	-1.639
V102	2	-0.713	-0.138
V303	1	-1.414	-1.639
V303	2	-0.787	-0.138
V303	3	-0.311	2.231
V103	1	-0.516	-1.639
V103	2	0.765	-0.138
V103	3	3.218	2.231
V304	1	-0.530	-1.639
V304	2	0.860	-0.138
V304	3	1.383	2.231
V304	4	3.616	5.466
V104	1	-2.288	-1.639
V104	2	-2.981	-1.639
V104	3	0.013	-1.639
V104	4	-0.552	-0.138

Item Set 10 - 12. LRT was not significant for the imputed data (p = .57). However, Martin-Löf-test was significant (p = .04). IFA was significant as well (p < .05). Item difficulty parameter correlated highly (r = .96, p < .001).

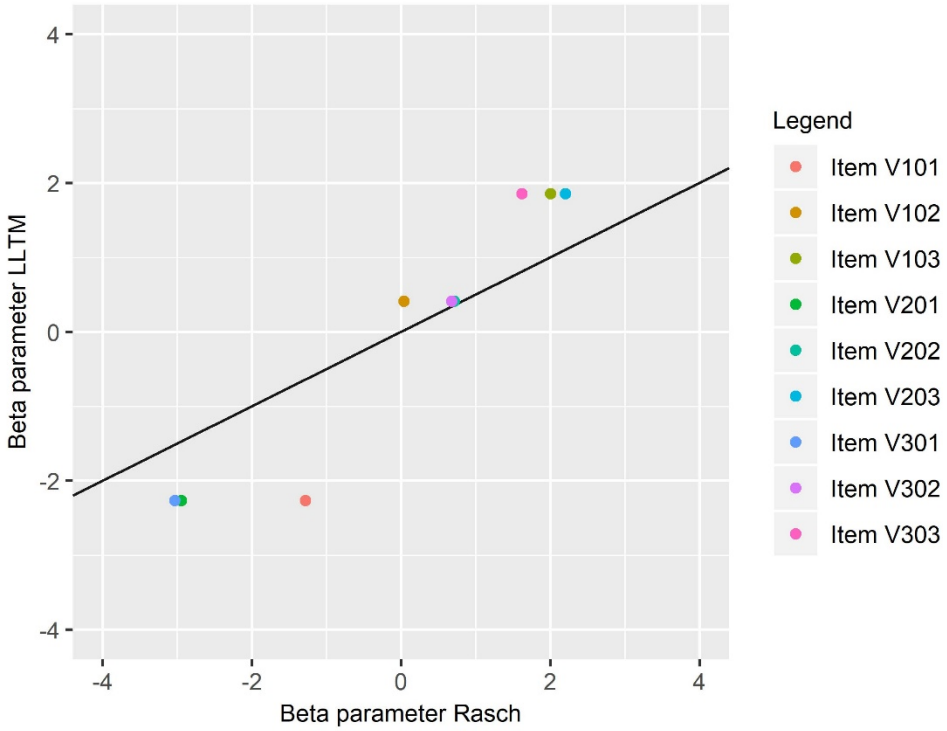


Figure 39. RM and LLTM beta parameter of Item Set 10 – 12.

Table 93

Item difficulty parameter for the RM and the LLTM - Item Set 10-12

Item	Item difficulty parameter Rasch	Item difficulty parameter LLTM
V101	-2.102	-3.420
V201	-4.081	-3.420
V301	-3.848	-3.420
V102	-0.590	-0.567
V202	-0.024	-0.567
,V302	-0.089	-0.567
V103	1.333	0.871
V203	1.672	0.871
V303	0.780	0.871
V104	3.615	3.117
V304	3.333	3.117

For the PCM, LRT showed no significance (p = .07) as well as Martin-Löf-test (p = .53). Item difficulty parameters of PCM and LPCM correlated with r = .86 (p < .001). Item difficulty parameters for the PCM and the LPCM without an a priori defined q-matrix correlated with r = 1 (p < .001).

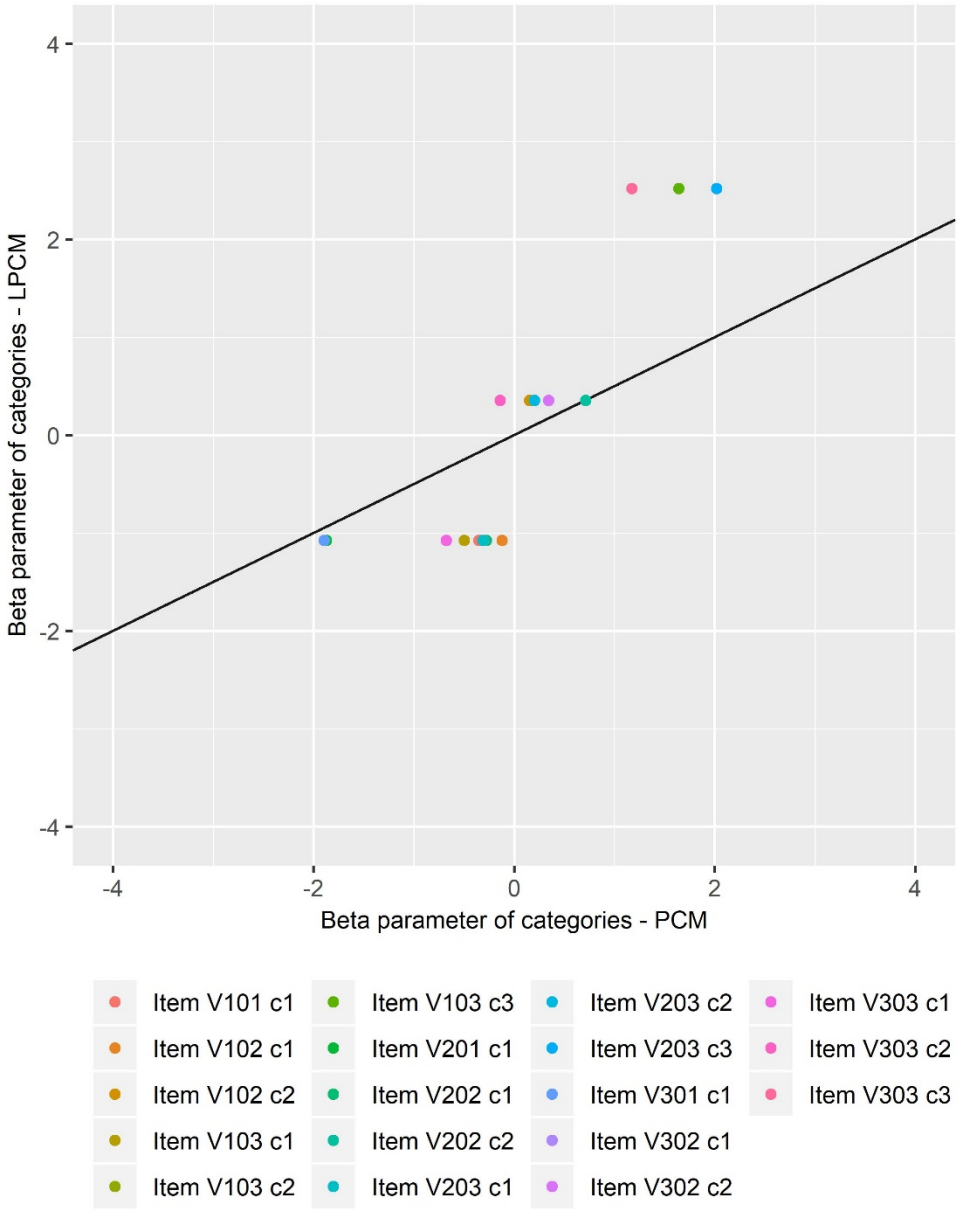


Figure 40. PCM and LPCM beta parameter of Item Set 10 – 12.

In a simulation, random q-matrices with different ratios of 0's and 1's were generated and the LPCM calculated with those. The item difficulty parameter of the PCM and the newly calculated LPCM were correlated and the minimal correlation, the median, the mean, the 95th percentile and the maximum correlation determined for each imputed dataset as can be seen in Table 94 to Table 103. Missing values could not be calculated due to the properties of the artificially generated design matrix.

Table 94

Descriptive statistics for the correlations obtained from simulated weight matrices for 20% occupancy – Item Set 10 - 12

$\%_1$	Min	Median	Mean	95%	Max
20	-.5719	.1131	.0888	.5478	.6305
20	-.2937	.1176	.1291	.4811	.7519
20	-.4488	.0700	.0811	.4232	.6728
20	-.4733	.0867	.1056	.5275	.7064
20	-.5320	.0838	.0731	.4759	.6718

Note. $\%_1$ = occupancy with 1's in the weight matrix; Min = minimum correlation, 95% = 95th percentile; Max = maximum correlation.

Table 95

Descriptive statistics for the correlations obtained from simulated weight matrices for 25% occupancy – Item Set 10 - 12

$\%_1$	Min	Median	Mean	95%	Max
25	-.4409	.1049	.1056	.4547	.7460
25	-.5636	.0941	.0632	.3864	.6064
25	-.4370	.0993	.0757	.3938	.6033
25	-.5134	.0178	.0564	.4823	.5834
25	-.4895	.1301	.1226	.4828	.6536

Note. $\%_1$ = occupancy with 1's in the weight matrix; Min = minimum correlation, 95% = 95th percentile; Max = maximum correlation.

Table 96

Descriptive statistics for the correlations obtained from simulated weight matrices for 30% occupancy – Item Set 10 - 12

$\%_1$	Min	Median	Mean	95%	Max
30	-.4820	.1373	.1187	.4772	.6160
30	-.5037	.0647	.0878	.5121	.6743
30	-.3956	.1308	.1510	.4828	.6555
30	-.4337	.0489	.0519	.4379	.8656
30	-.4116	.1210	.1174	.4809	.6268

Note. $\%_1$ = occupancy with 1's in the weight matrix; Min = minimum correlation, 95% = 95th percentile; Max = maximum correlation.

Table 97

Descriptive statistics for the correlations obtained from simulated weight matrices for 35% occupancy – Item Set 10 - 12

$\%_1$	Min	Median	Mean	95%	Max
35	-.4787	.1478	.1280	.4437	.6813
35	-.5412	.1200	.1083	.4900	.8240
35	-.3996	.1207	.1289	.5070	.6494
35	-.5366	.0913	.1095	.4962	.6655
35	-.4224	.1383	.1244	.5391	.5835

Note. $\%_1$ = occupancy with 1's in the weight matrix; Min = minimum correlation, 95% = 95th percentile; Max = maximum correlation.

Table 98

Descriptive statistics for the correlations obtained from simulated weight matrices for 40% occupancy – Item Set 10 - 12

$\%_1$	Min	Median	Mean	95%	Max
40	-.5096	.1297	.1209	.4614	.5394
40	-.5194	.1401	.1338	.5024	.5892
40	-.2785	.1643	.1450	.4301	.6282
40	-.3816	.1109	.1228	.4889	.5547
40	-.3221	.1103	.1233	.5315	.7581

Note. $\%_1$ = occupancy with 1's in the weight matrix; Min = minimum correlation, 95% = 95th percentile; Max = maximum correlation.

Table 99

Descriptive statistics for the correlations obtained from simulated weight matrices for 45% occupancy – Item Set 10 - 12

$\%_1$	Min	Median	Mean	95%	Max
45	-.2871	.1414	.1306	.4247	.5611
45	-.4451	.1570	.1531	.5602	.7044
45	-.3887	.1748	.1806	.5530	.6532
45	-.3724	.1779	.1504	.4877	.5977
45	-.5209	.1002	.0968	.3960	.5686

Note. $\%_1$ = occupancy with 1's in the weight matrix; Min = minimum correlation, 95% = 95th percentile; Max = maximum correlation.

Table 100

Descriptive statistics for the correlations obtained from simulated weight matrices for 50% occupancy – Item Set 10 - 12

$\%_1$	Min	Median	Mean	95%	Max
50	-.3486	.1885	.1743	.5204	.7549
50	-.4181	.1499	.1469	.5178	.6257
50	-.2606	.1920	.2010	.5409	.6621
50	-.3152	.1447	.1352	.4401	.5755
50	-.4280	.1711	.1834	.5721	.6592

Note. $\%_1$ = occupancy with 1's in the weight matrix; Min = minimum correlation, 95% = 95th percentile; Max = maximum correlation.

Table 101

Descriptive statistics for the correlations obtained from simulated weight matrices for 55% occupancy – Item Set 10 - 12

$\%_1$	Min	Median	Mean	95%	Max
55	-.5110	.1701	.1649	.4874	.6492
55	-.5424	.1604	.1653	.4819	.6001
55	-.3536	.1413	.1505	.5056	.6079
55	-.3727	.2028	.2098	.5532	.6662
55	-.3370	.1190	.1504	.5220	.6809

Note. $\%_1$ = occupancy with 1's in the weight matrix; Min = minimum correlation, 95% = 95th percentile; Max = maximum correlation.

Table 102

Descriptive statistics for the correlations obtained from simulated weight matrices for 60% occupancy – Item Set 10 - 12

$\%_1$	Min	Median	Mean	95%	Max
60	-.3398	.1568	.1637	.5280	.6811
60	-.3642	.1984	.1756	.5490	.6403
60	-.5593	.2211	.2045	.5570	.6428
60	-.2752	.1850	.2113	.6173	.6726
60	-.2984	.2500	.2309	.5503	.6209

Note. $\%_1$ = occupancy with 1's in the weight matrix; Min = minimum correlation, 95% = 95th percentile; Max = maximum correlation.

Table 103

Descriptive statistics for the correlations obtained from simulated weight matrices for 65% occupancy – Item Set 10 - 12

$\%_1$	Min	Median	Mean	95%	Max
65	-.1458	.2293	.2241	.5743	.7034
65	-.2815	.2589	.2355	.5538	.6447
65	-.2313	.1761	.1937	.5579	.7719
65	-.2724	.2230	.2443	.6154	.8763
65	-.1942	.2193	.2216	.5377	.7511

Note. $\%_1$ = occupancy with 1's in the weight matrix; Min = minimum correlation, 95% = 95th percentile; Max = maximum correlation.

In a second simulation, q-matrices were permutated. The results of the descriptive statistics for each imputed dataset can be seen in Table 104.

Table 104

Descriptive statistics for the correlations obtained from permutated simulated weight matrices – Item Set 10 - 12

Min	Median	Mean	95%	Max
.2461	.7310	.6553	.8608	.8608
.2305	.6836	.6249	.8183	.8183
.2430	.7352	.6487	.8537	.8537
.2135	.7245	.6317	.8386	.8386
.2760	.7607	.6759	.8677	.8677

Note. Min = minimum correlation, 95% = 95th percentile; Max = maximum correlation.

Table 105

Item difficulty parameter for the PCM and the LPCM - Item Set 10 - 12

Item	Category	Item difficulty parameter PCM	Item difficulty parameter LPCM
V101	1	-0.353	-1.075
V201	1	-1.873	-1.075
V301	1	-1.893	-1.075
V101	1	-0.121	-1.075
V102	2	0.152	0.353
V201	1	-0.275	-1.075
V202	2	0.713	0.353
V301	1	-0.306	-1.075
V302	2	0.343	0.353
V101	1	-0.498	-1.075
V102	2	0.198	0.353
V103	3	1.643	2.518
V201	1	-0.311	-1.075
V202	2	0.205	0.353
V203	3	2.020	2.518
V301	1	-0.676	-1.075
V302	2	-0.139	0.353
V303	3	1.174	2.518

Simulation study: Sample size

Background. The question of sample size and associated issues are hotly debated when it comes to RM (e.g., Babcock & Hodge, 2020). Different approaches have been applied to determine the ideal sample size (e.g., Draxler & Alexandrowicz, 2015). Baker (1993) and MacDonald (2014) pointed out that a minimum sample size of $N = 50$ is necessary for appropriate parameter estimation. However, this should be checked again under the conditions prevailing in this study using a dense q-matrix.

Current study. The aim of this simulation study was to determine the sample size at which parameter estimation becomes acceptable given the existing item sets of Study 1. In addition, to what extent the number of items affects the results should be assessed.

Methods. For this purpose, data were simulated under different conditions. The sample size was systematically varied (40, 45, 50, 100, 250, 500, 1000, 2500). The small steps at the lower end of the spectrum can be attributed to the fact that Baker (1993) and MacDonald (2014) indicate that parameter estimation becomes stable beginning at $N = 50$; therefore, smaller sample sizes were deliberately tested. In addition, a minimum of 6 items was assumed. Then, analogously to the item sets used in Study 1, a maximum of 12 items was reached with a sequence of two. Randomized beta parameters were created, although two beta parameters in one dataset were always identical. This is analogous to an optimal dataset in Study 1. The beta parameters were limited to a range of -4 to 3. In addition, thetas were created with $M = 0$ and $SD = 1.5$. Based on these values, simulated answeres to the items were created and corresponding beta parameters were calculated using an LLTM. The beta parameters determined by the LLTM were then correlated with the original beta

parameters. The replication rate was $r = 2000$. In the end, the minimum, mean, median, 95% quantile and maximum correlation were calculated for each condition.

Results. All results can be found in Table 106.

Table 106

Correlation coefficients between original beta parameter and beta parameter obtained from the simulated response patterns with r = 2000

N	*k*	Min	Median	Mean	95%	Max
40	6	-.96	1.00	.97	1.00	1.00
45		-.93	1.00	.98	1.00	1.00
50		-.82	1.00	.97	1.00	1.00
100		-.99	1.00	.99	1.00	1.00
250		-.39	1.00	.99	1.00	1.00
500		.31	1.00	1.00	1.00	1.00
1000		.18	1.00	1.00	1.00	1.00
2500		.72	1.00	1.00	1.00	1.00
40	8	-.92	.99	.98	1.00	1.00
45		-.41	.99	.98	1.00	1.00
50		-.16	.99	.99	1.00	1.00
100		.34	1.00	.99	1.00	1.00
250		.76	1.00	1.00	1.00	1.00
500		-.46	1.00	1.00	1.00	1.00
1000		.86	1.00	1.00	1.00	1.00
2500		.93	1.00	1.00	1.00	1.00
40	10	.58	.99	.98	.99	1.00
45		.41	.99	.99	.99	1.00

continued

continued

N	*k*	Min	Median	Mean	95%	Max
50	10	.02	.94	.91	.99	1.00
100		.25	.94	.91	1.00	1.00
250		.64	.95	.91	1.00	1.00
500		.27	.95	.91	1.00	1.00
1000		.21	.95	.91	1.00	1.00
2500		.21	.95	.91	1.00	1.00
40	12	.38	.99	.98	1.00	1.00
45		.66	.99	.99	1.00	1.00
50		.85	.99	.98	1.00	1.00
100		.86	1.00	1.00	1.00	1.00
250		.93	1.00	1.00	1.00	1.00
500		.97	1.00	1.00	1.00	1.00
1000		.99	1.00	1.00	1.00	1.00
2500		1.00	1.00	1.00	1.00	1.00

Note: *k* = number of items, Min = minimum correlation, 95% = 95th percentile; Max = maximum *p*-value.

Discussion. The aim of the present simulation study was to determine the sample size at which parameter estimation is stable enough under the conditions of Study 1. It was shown that although more than half the correlation coefficients were very good, with ≥ .98, there are outliers in the minimum values, even at larger sample sizes. This problem is even more serious when the number of items is small. The best results are achieved with 12 items and a minimum sample size of $N = 250$. Accordingly, a minimum sample size of $N = 250$ can be recommended when calibrating 12 items.

Simulation study: Recovery of beta parameters via multiple imputation and predictive mean matching

Background. IFA (Bock et al., 1988) is a special case of structural equation modeling (MacCallum, 2009) that aims to evaluate dimensionality (e.g., Mair, 2018). Therefore, it can be applied to determine violations of the RM assumptions. Due to problems with the IFA (the two-factor IFA model was always superior to the one factor-model) in Study 1 with the imputed data recovered through predictive mean matching, a Monte-Carlo simulation study was conducted.

Current study. The study had three aims: the first aim was to investigate whether parameter and response pattern recovery was sufficient to perform an IFA. Specifically, it was tested whether the IFA accurately detects unidimensionality with the recovered parameters.

The second aim was to determine whether parameter recovery was sufficient, since multiple imputation seems to be problematic in the present case.

The third aim was to determine whether the data structure caused the significant IFA.

Methods.

Study A. In a first step, the impact of the imputed data on IFA was investigated. Different simulations were conducted for this purpose. To replicate the multiple imputation conditions described in Study 1 (Item Sets 1-3, 4-6, 7-9, 10-12), beta parameters were generated for each fictitious item set (3x6, 18 in total), which reflected the number of items in the imputed dataset in Study 1. Different response patterns were generated in terms of unidimensional thetas ($M = 0$, $SD = 1.5$). Two conditions were examined: one with $N = 10{,}002$ and one with a smaller sample ($N = 120$). This was important for two reasons: first, a large sample should not show any effects due to the range of ability parameters and response patterns. Second, a smaller sample mimics the conditions in Study 1. So that data could be deleted later on, similar to the missing data in Study 1, the number of simulated participants needed to be divisible by three. The replication rate was set to $r = 2{,}000$ for each simulation. Responses were simulated using the sim.rasch function from the eRm package (Mair et al., 2019).

In a first step, two simulations were conducted for each condition as a proof of concept: First, randomized betas between -4 and 3 were drawn, similar to the range of beta parameters for the imputed data in Study 1. Subsequently, the betas were fixed and assumed equal within a specific item cluster, as is the case in a LLTM. The range was chosen to be similar to the beta parameter in the first study and should reflect the beta parameters for a LLTM. Therefore, beta parameters were set to -4, -4, -4, -3, -3, -3, -1, -1, -1, 0, 0, 0, 1, 1, 1, 3, 3 and 3 (for an example R script, see the addendum to this study). In all cases, IFAs for a one-factor- and two-factor-solution were computed and compared via ANOVA. The resulting *p*-values were stored. At the end of the simulation, the minimum, maximum, mean, median and 95% quantile *p*-values were calculated.

The second step followed an equivalent procedure to the first step; however, data was missing. To replicate the conditions of Study 1, a third of responses were removed. Of these missing responses, one-third involved removing responses to the fictitious Items 1, 4, 7, 10, 13 and 16; another third responses to Items 2, 5, 8, 11, 14 and 17; and the final third responses to Items 3, 6, 9, 12, 15 and 18. As in Study 1, multiple imputation with predictive mean matching (pmm) was conducted with 5 iterations. All other operations were identical as in Step 1.

Study B. This study focused not only on IFA, but also on the recovery of beta parameters. For that purpose, the procedure from Study A was repeated, but a LLTM rather than an IFA was calculated for each dataset. The beta parameters were standardized and correlated with the original beta parameters set at the beginning of the simulation. The median, mean, minimum, maximum and 95th percentile of the correlation coefficients were obtained.

Study C. The final study sought to determine whether the data structure of the LLTM causes the problem. Therefore, data were simulated using four equal beta parameters on each level, enforcing the LLTM data structure. Beta parameters were randomly drawn within a range of -4 and 3, replicating the conditions of Study 1. Different response patterns were generated in the form of unidimensional thetas ($M = 0$, $SD = 1.5$). Subsequents, IFAs for a one-factor model and a two-factor model were computed and compared via ANOVA. The median, mean, minimum, maximum and 95th percentile of the *p*-values were obtained. Furthermore, different sample sizes (50, 100, 200, 500 and 1,000) were simulated. The replication rate was $r = 2000$.

Results.

Study A. All descriptive statistics for the IFA *p*-values of the IFA can be seen in Table 107.

Table 107

Descriptive statistics for IFA p-values obtained from simulated response patterns with r = 2000

N	*β*	Missing values	Min	Median	Mean	95%	Max
120	random	no	.00	.04	.08	.26	.64
10,002	random	no	.00	.05	.09	.31	.65
120	fixed	no	.00	.04	.07	.25	.69
10,002	fixed	no	.00	.05	.08	.30	.69
120	random	yes	.00	.00	.00	.00	.00
10,002	random	yes	.00	.00	.00	.00	.00
120	fixed	yes	.00	.00	.00	.00	.00
10,002	fixed	yes	.00	.00	.00	.00	.00

Note: Min = minimum correlation, 95% = 95th percentile; Max = maximum *p*-value.

Study B. All correlation coefficients and associated descriptive statistics can be seen in Table 108.

Table 108

Descriptive statistics for the correlations obtained from simulated response patterns with r = 2000

N	*β*	Missing values	Min	Median	Mean	95%	Max
120	random	no	.11	.51	.51	.75	.94
10,002	random	no	.10	.50	.51	.77	.95
120	fixed	no	.99	1.00	1.00	1.00	1.00
10,002	fixed	no	1,00	1.00	1.00	1.00	1.00
120	random	yes	.05	.52	.52	.74	.94
10,002	random	yes	.06	.52	.52	.75	.93
120	fixed	yes	.98	1.00	1.00	1.00	1.00
10,002	fixed	yes	1.00	1.00	1.00	1.00	1.00

Note: Min = minimum correlation, 95% = 95th percentile; Max = maximum correlation.

Study C. All descriptive statistics for the IFA *p*-values can be seen in Table 109.

Table 109

Descriptive statistics for the IFA p-values obtained from simulated response patterns with r = 2000

N	Min	Median	Mean	95%	Max
50	.00	.01	.03	.13	.41
100	.00	.02	.04	.16	.42
200	.00	.02	.05	.18	.56
500	.00	.02	.04	.17	.43
1000	.00	.02	.05	.15	.55

Note: Min = minimum correlation, 95% = 95th percentile; Max = maximum *p*-value.

Discussion. As shown in Study A, IFA seems to be relatively reliable in detecting violations of unidimensionality. In both the small sample condition with $N = 102$ and the large sample condition with $N = 10{,}002$, the mean significance level is above $p = .05$. This applies to both random beta parameters and fixed beta parameters. However, IFAs were always significant as soon as imputed missing values appeared in the dataset. There seem to have been overall problems with the IFA, since the median and mean *p*-values were close to significance level even when no data was missing. In the present case of multiple imputation, IFA has no validity with respect to unidimensionality, since the IFA was always significant even though values were deleted from a unidimensional dataset.

Since problems with IFA were already foreseeable in Study A, Study B additionally investigated whether the beta parameters were correctly reconstructed by multiple imputation. All imputed datasets had very high correlations with the preset beta parameters of the original datasets, suggesting that the reconstructed response patterns still seem to be sufficient. As expected, the simulations with random item parameters led to weaker correlation coefficients, since the beta parameters did not fit the design matrix. However, the results for the complete and imputed datasets were similar, further that multiple imputation indicating recovers this parameter sufficiently.

Overall, multiple imputation using pmm seems to lead to problems with IFA, but does not seem to have a significant effect on parameter estimation.

However, Study C could show that the IFA seems to have problems with the data structure, because it became significant even at large sample sizes. Therefore, the IFA seems to be unsuited for detecting violations of assumptions under the present conditions.

Addendum.

Example for a R Script – Study A.

```
gc()
rm(list=ls())

library(here)
library(eRm)
library(psych)
library(Hmisc)
library(mice)
library(mirt)

replicationrate <- 100
tmp <- matrix(NA, nrow = replicationrate, ncol=1)
startzeit <- Sys.time()

for (j in 1:replicationrate){

 no_theta <- 10002 #muss durch 3 teilbar sein
 theta <- rnorm(no_theta,0,1.5) # Thetas erstellen, sollte durch 3 teilbar sein
 beta <- runif(18,min=-4, max=3) # Betas erstellen für 18 Items
 rand <- sample(1:999999999,1) # randomisierte Zahl zur Simulation
 data_comp<- eRm::sim.rasch(theta, beta, seed=rand)

 dat <- data_comp
 for (i in 1:nrow(dat)){
  dat[i, seq(i%%3, 18, by=3)] <- NA # jede dritte Zahl wird im dataframe durch
  " NA" ersetzt
 }
 MAXITER <- 5
 ran2 <- sample(1:999999999,1) # randomisierte Zahl zur Simulation
```

```
imp <- mice(dat,m=MAXITER, method = "pmm" ,seed=ran2, printFlag = F)

no_items <- length(beta)
tmpL <- matrix(NA, nrow = MAXITER, ncol=1) # ncol Nummmer der Items

for (i in 1:MAXITER){
 Sim_voll <- complete(imp, i)
 Ident <- seq(1,18,1)
 colnames(Sim_voll) <- Ident
 i_Modell1 <- tryCatch(mirt(Sim_voll,1,verbose=F), error=function(e)
 return(NA) )
 i_Modell2 <- tryCatch(mirt(Sim_voll,2,verbose=F), error=function(e)
 return(NA) )
 comp_M <- tryCatch(anova(i_Modell1,i_Modell2), error=function(e)
 return(list(p=c(NA,NA) )))
 tmpL[i, 1] <- comp_M$p[2]
}
# wenn auf Grund der Imputation nicht gerechnet werden kann, weil eine Variable
entfernt wird, wird hier in den Vektor "NA" eingetragen

mean.p <- colMeans(tmpL, na.rm=T)  #Berechnung der Mw für alle p-s
tmp[j, 1] <- mean.p

## Zeitschleife zur Angabe der Zeit
if (j %% 10 ==0){ # Alle 10 Iterationen wird Zeit berechnet
 zeit <- Sys.time()
 differenz <- difftime(zeit, startzeit, units=c('secs'))
 print(paste(j, "Iterationen fertig. Dauer: ", round(differenz, 1), " Sekunden"))
 rest <- differenz / j * replicationrate - differenz
 print(paste('Verbleibende Restzeit ca. ', round(rest, 1), ' Sekunden'))
 print(paste('Das sind ca. ', round(rest/60, 1), ' Minuten'))
}}
```

Example for a R Script – Study B.

```
gc()
rm(list=ls())

library(here)
library(eRm)
library(psych)
library(Hmisc)
library(mice)

replicationrate <- 2000
tmp <- matrix(NA, nrow = replicationrate, ncol=1)

startzeit <- Sys.time()

for (j in 1:replicationrate){
 no_theta <- 120 #muss durch 3 teilbar sein
 theta <- rnorm(no_theta,0,1.5) # Thetas erstellen, sollte durch 3 teilbar sein
 beta <- c(-4,-4,-4,-3,-3,-3,-1,-1,-1, 0,0,0,1,1,1,3,3,3) # Betas erstellen für 18 Items
 rand <- sample(1:999999999,1) # randomisierte Zahl zur Simulation
 data_comp<- eRm::sim.rasch(theta, beta, seed=rand)
 Q1=matrix(c(
  0,0,0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
  0,0,0,0,0,0,1,1,1,1,1,1,1,1,1,1,1,1,
  0,0,0,0,0,0,0,0,0,1,1,1,1,1,1,1,1,1,
  0,0,0,0,0,0,0,0,0,0,0,0,1,1,1,1,1,1,
  0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,1),ncol =5) #hier Anzahl der Items
 dat <- data_comp
 for (i in 1:nrow(dat)){
  dat[i, seq(i%%3, 18, by=3)] <- NA # jede dritte Zahl wird im dataframe durch
   "NA" ersetzt
 }
```

```
MAXITER <- 5
ran2 <- sample(1:999999999,1) # randomisierte Zahl zur Simulation
imp <- mice(dat,m=MAXITER, method = "pmm" ,seed=ran2, printFlag = F)
no_items <- length(beta)
tmpL <- matrix(NA, nrow = MAXITER, ncol=no_items) # ncol Nummmer der
Items
for (i in 1:MAXITER){
  Sim_voll <- complete(imp, i)
  i_Modell <- tryCatch(LLTM(Sim_voll,Q1), error=function(e)
  return(list(betapar=rep(NA, no_items))) )
  # wenn auf Grund der Imputation nicht gerechnet werden kann, weil eine
  Variable entfernt wird, wird hier in den Vektor n_col Mal "NA" eingetragen
  tmpL[i, 1:length(i_Modell$betapar)] <- i_Modell$betapar
}
betaparam.L <- colMeans(tmpL, na.rm=T)  #Berechnung der Mw für alle
 Itemparameter
beta.L <- ((round(betaparam.L,3)-round(mean(betaparam.L),3))*(-1))

## Zusammenhang Betaparameter
Zsmhang <- rcorr(beta,beta.L)
Korrel <- Zsmhang$r[2,1]

tmp[j, 1] <- Korrel

## Zeitschleife zur Angabe der Zeit
if (j %% 10 ==0){ # Alle 10 Iterationen wird Zeit berechnet
  zeit <- Sys.time()
  differenz <- difftime(zeit, startzeit, units=c('secs'))
  print(paste(j, "Iterationen fertig. Dauer: ", round(differenz, 1), " Sekunden"))
  rest <- differenz / j * replicationrate - differenz
  print(paste('Verbleibende Restzeit ca. ', round(rest, 1), ' Sekunden'))
  print(paste('Das sind ca. ', round(rest/60, 1), ' Minuten'))}}
```

Detailed results Study 2

Model fit.

Working memory figural. LRT (p = .24), Martin-Löf-test (p = .12) and Waldtest showed no significance (p > .05). The χ^2/df of the LRT was 1.20. Item difficulty parameter correlated with r = .98 (p < .001).

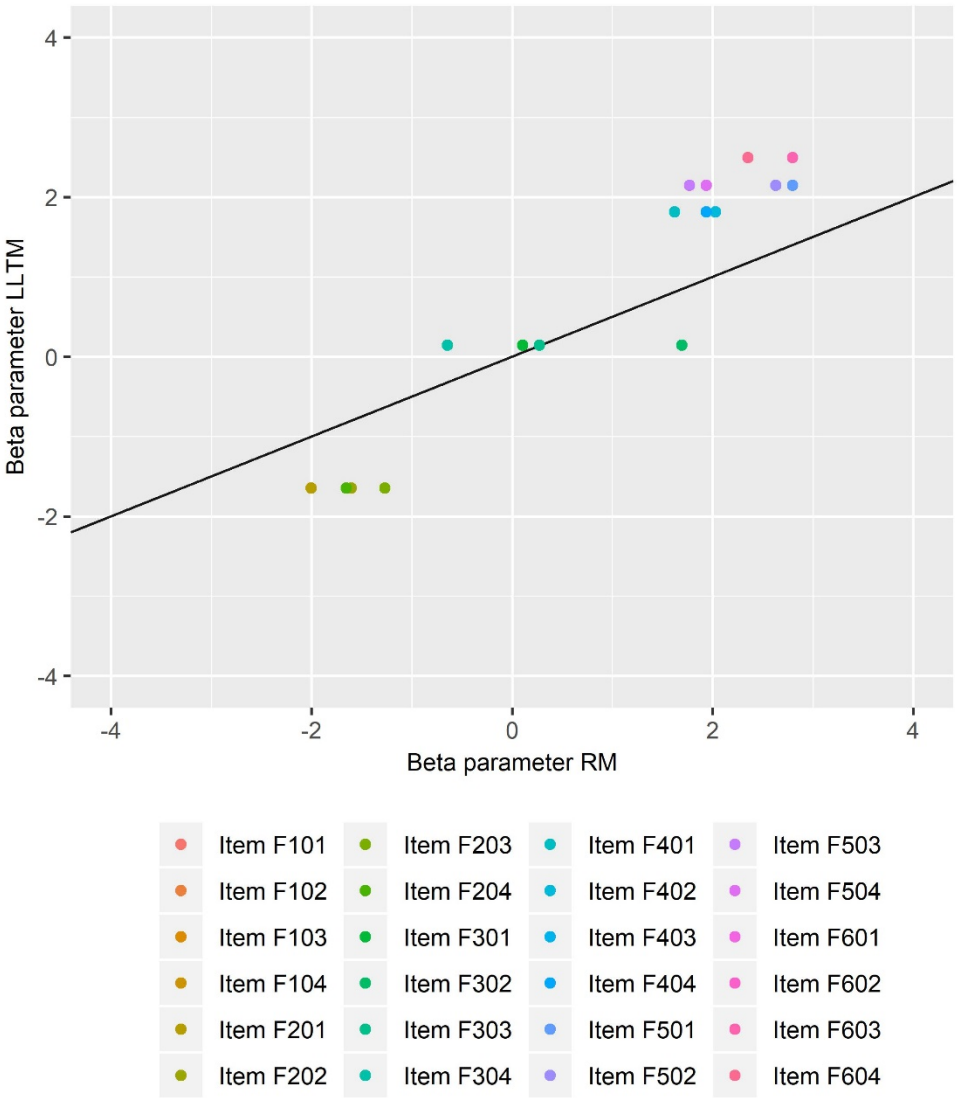


Figure 41. RM and LLTM beta parameter of WM-F.

Table 110

Item difficulty parameter for the RM and the LLTM – WM-F

Item	Item difficulty parameter Rasch	Item difficulty parameter LLTM
F101	-5.029	-4.966
F201	-5.958	-4.966
F301	-6.914	-4.966
F401	-3.914	-4.966
F102	-2.006	-1.643
F202	-1.605	-1.643
F302	-1.271	-1.643
F402	-1.655	-1.643
F103	0.106	0.148
F203	1.692	0.148
F303	0.270	0.148
F403	-0.646	0.148
F104	1.621	1.817
F204	2.026	1.817
F304	1.934	1.817
F404	1.934	1.817
F105	2.796	2.150
F205	2.628	2.150
F305	1.768	2.150
F405	1.934	2.150
F106	2.350	2.494
F206	2.796	2.494
F306	2.796	2.494
F406	2.350	2.494

For the PCM, LRT showed no significance ($p = .42$) as well as Martin-Löf-test ($p = 1.0$). Item difficulty parameters of PCM and LPCM correlated with $r = .86$ ($p < .001$).

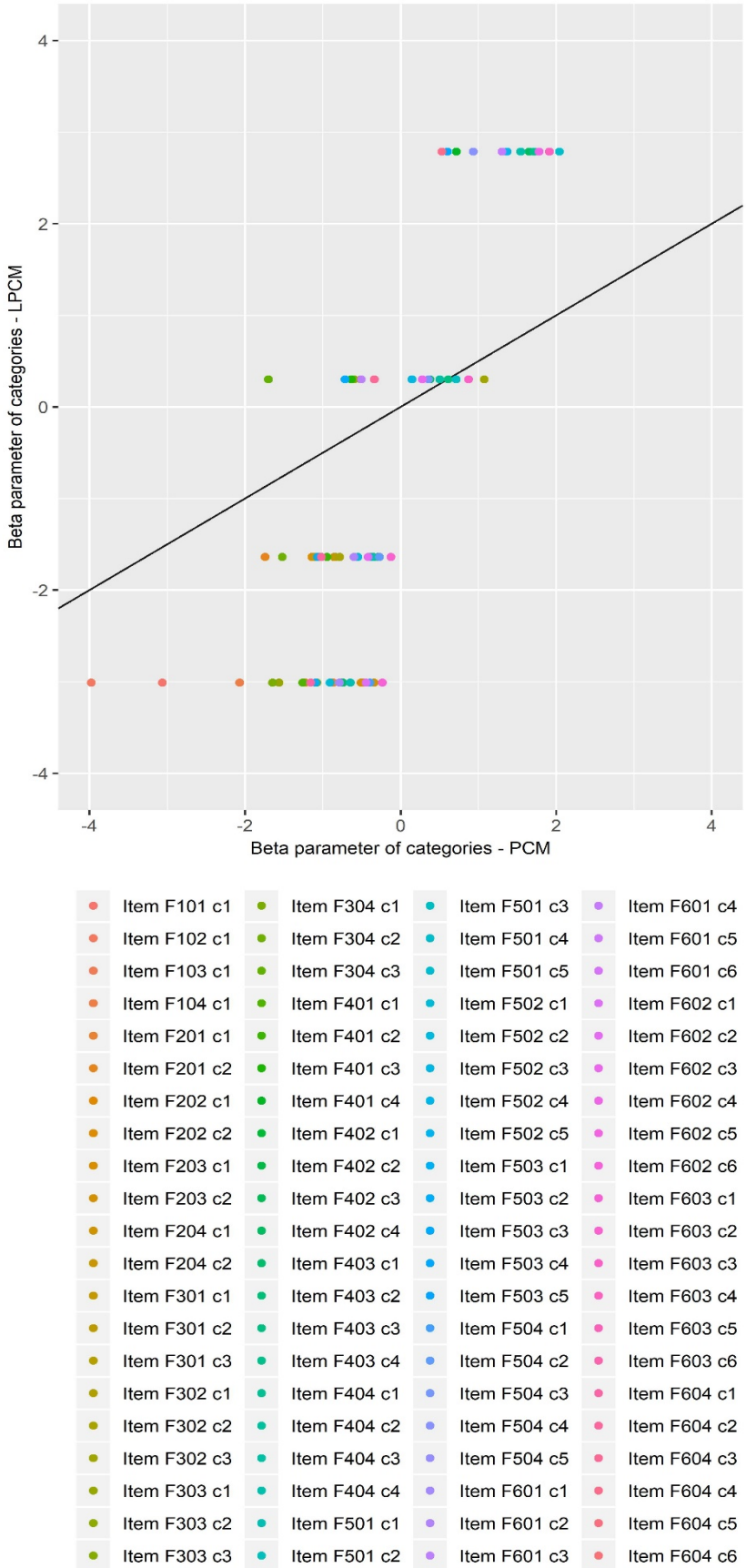


Figure 42. PCM and LPCM beta parameter of categories of WM-F.

In a simulation, random q-matrices with different ratios of 0's and 1's were generated and the LPCM calculated with those. The item difficulty parameter of the PCM and the newly calculated LPCM were correlated and the minimal correlation, the median, the mean, the 95th percentile and the maximum correlation determined as can be seen in Table 111.

Table 111

Descriptive statistics for the correlations obtained from simulated weight matrices – WM-F

$\%_1$	Min	Median	Mean	95%	Max
20	-.2609	.0088	.0148	.1897	.2410
25	-.2897	.0042	.0073	.1604	.2138
30	-.2264	.0109	.0158	.1898	.3185
35	-.2436	.0167	.0193	.1813	.2539
40	-.2402	.0192	.0214	.1894	.2415
55	-.2687	.0190	.0114	.1589	.2611
60	-.2166	.0395	.0318	.2217	.2648
65	-.2816	.0319	.0352	.2174	.3100
70	-.3009	.0334	.0403	.2255	.3011

Note: $\%_1$ = occupancy with 1's in the weight matrix; Min = minimum correlation, 95% = 95th percentile; Max = maximum correlation.

In a second simulation, q-matrices were permutated. The results of the descriptive statistics can be seen in Table 112.

Table 112

Descriptive statistics for the correlations obtained from permutated simulated weight matrices – WM-F

Min	Median	Mean	95%	Max
.5696	.7907	.7599	.8663	.8685

Note: Min = minimum correlation, 95% = 95th percentile; Max = maximum correlation.

Table 113

Item difficulty parameter for the PCM and the LPCM – WM-F

Item	Category	Item difficulty parameter PCM	Item difficulty parameter LPCM
F101	1	-3.064	-3.010
F201	1	-3.977	-3.010
F301	1	-4.932	-3.010
F401	1	-2.070	-3.010
F101	1	-0.865	-3.010
F102	2	-1.743	-1.639
F201	1	-0.479	-3.010
F202	2	-1.142	-1.639
F301	1	-0.509	-3.010
F302	2	-0.855	-1.639
F401	1	-0.341	-3.010
F402	2	-1.110	-1.639
F101	1	-1.223	-3.010
F102	2	-0.841	-1.639
F103	3	-0.514	0.300
F201	1	-1.074	-3.010
F202	2	-0.781	-1.639
F203	3	1.075	0.300

continued

continued

Item	Category	Item difficulty parameter PCM	Item difficulty parameter LPCM
F301	1	-1.564	-3.010
F302	2	-1.035	-1.639
F303	3	-0.601	0.300
F401	1	-1.646	-3.010
F402	2	-1.522	-1.639
F403	3	-1.702	0.300
F101	1	-1.257	-3.010
F102	2	-0.947	-1.639
F103	3	-0.636	0.300
F104	4	0.717	2.787
F201	1	-0.752	-3.010
F202	2	-0.599	-1.639
F203	3	0.380	0.300
F204	4	1.652	2.787
F301	1	-0.743	-3.010
F302	2	-0.287	-1.639
F303	3	0.609	0.300
F304	4	1.714	2.787
F401	1	-1.088	-3.010
F402	2	-0.351	-1.639
F403	3	0.504	0.300
F404	4	1.543	2.787
F101	1	-0.644	-3.010
F102	2	-0.393	-1.639
F103	3	0.713	0.300
F104	4	2.042	2.787
F105	5	2.958	4.979
F201	1	-0.910	-3.010

continued

continued

Item	Category	Item difficulty parameter PCM	Item difficulty parameter LPCM
F202	2	-0.550	-1.639
F203	3	0.144	0.300
F204	4	1.367	2.787
F205	5	2.448	4.979
F301	1	-1.094	-3.010
F302	2	-1.069	-1.639
F303	3	-0.715	0.300
F304	4	0.603	2.787
F305	5	1.026	4.979
F401	1	-0.393	-3.010
F402	2	-0.274	-1.639
F403	3	0.360	0.300
F404	4	0.936	2.787
F405	5	1.887	4.979
F101	1	-0.787	-3.010
F102	2	-0.600	-1.639
F103	3	-0.503	0.300
F104	4	1.302	2.787
F105	5	2.650	4.979
F106	6	2.401	6.737
F201	1	-0.444	-3.010
F202	2	-0.415	-1.639
F203	3	0.279	0.300
F204	4	1.779	2.787
F205	5	2.513	4.979
F206	6	3.280	6.737
F301	1	-0.233	-3.010
F302	2	-0.125	-1.639

continued

continued

Item	Category	Item difficulty parameter PCM	Item difficulty parameter LPCM
F303	3	0.874	0.300
F304	4	1.913	2.787
F305	5	4.127	4.979
F306	6	3.699	6.737
F401	1	-1.155	-3.010
F402	2	-1.018	-1.639
F403	3	-0.338	0.300
F404	4	0.532	2.787
F405	5	1.798	4.979
F406	6	2.086	6.737

Working memory verbal. LRT (p = .73), Martin-Löf-test (p = 1.0) and Waldtest showed no significance ($p < .05$). The χ^2/df of the LRT was 0.70. Item difficulty parameter correlated with r = .97 ($p < .001$).

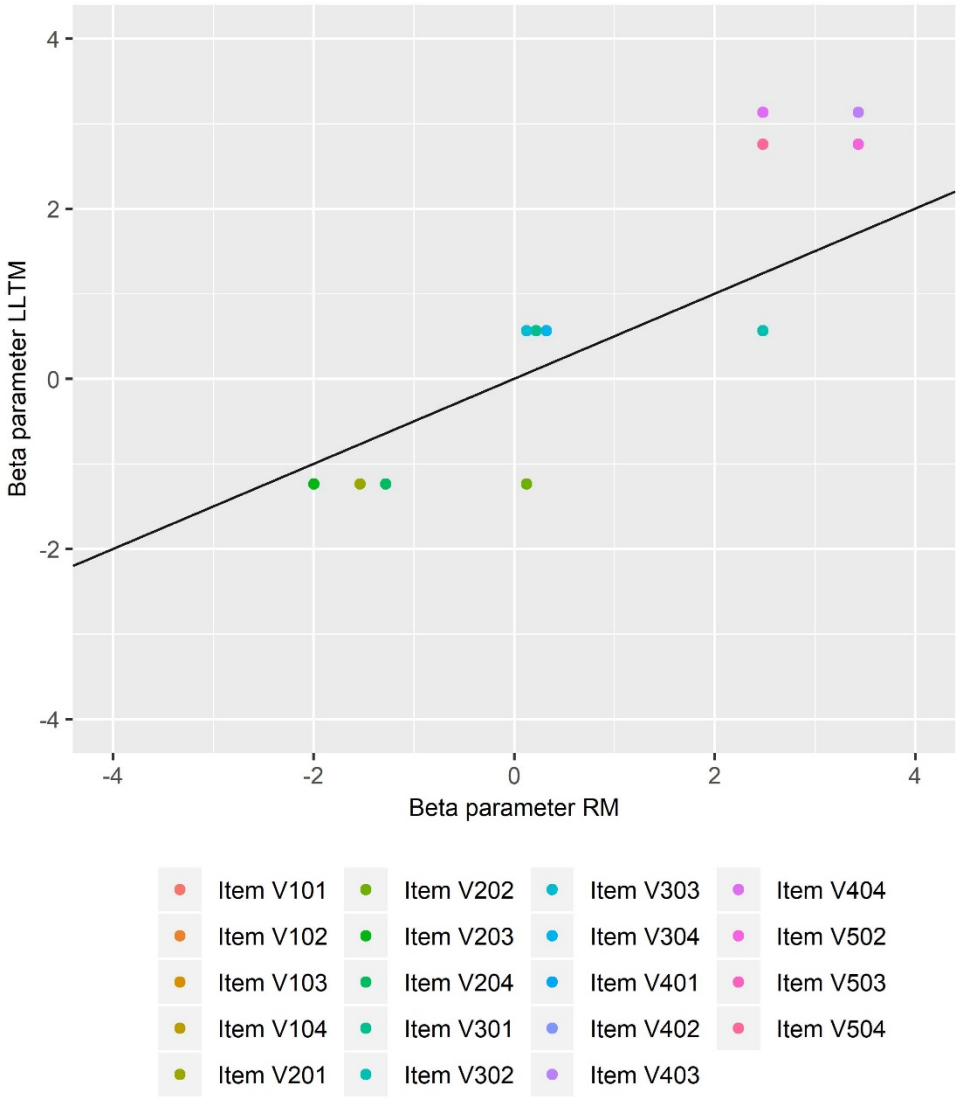


Figure 43. RM and LLTM beta parameter of WM-V.

Table 114

Item difficulty parameter for the RM and the LLTM – WM-V

Item	Item difficulty parameter Rasch	Item difficulty parameter LLTM
V101	-3.569	-4.535
V201	-4.908	-4.535
V301	-5.348	-4.535
V401	-5.780	-4.535
V102	-1.538	-1.234
V202	0.123	-1.234
V302	-1.998	-1.234
V402	-1.282	-1.234
V103	0.217	0.565
V203	2.479	0.565
V303	0.123	0.565
V403	0.320	0.565
V104	3.431	3.135
V204	3.431	3.135
V304	3.431	3.135
V404	2.479	3.135
V205	3.431	2.758
V305	2.479	2.758
V405	2.479	2.758

For the PCM, LRT showed significance ($p < .05$). However, Martin-Löf-test was not significant ($p = 1.0$). Item difficulty parameters of PCM and LPCM correlated with $r = .88$ ($p < .001$).

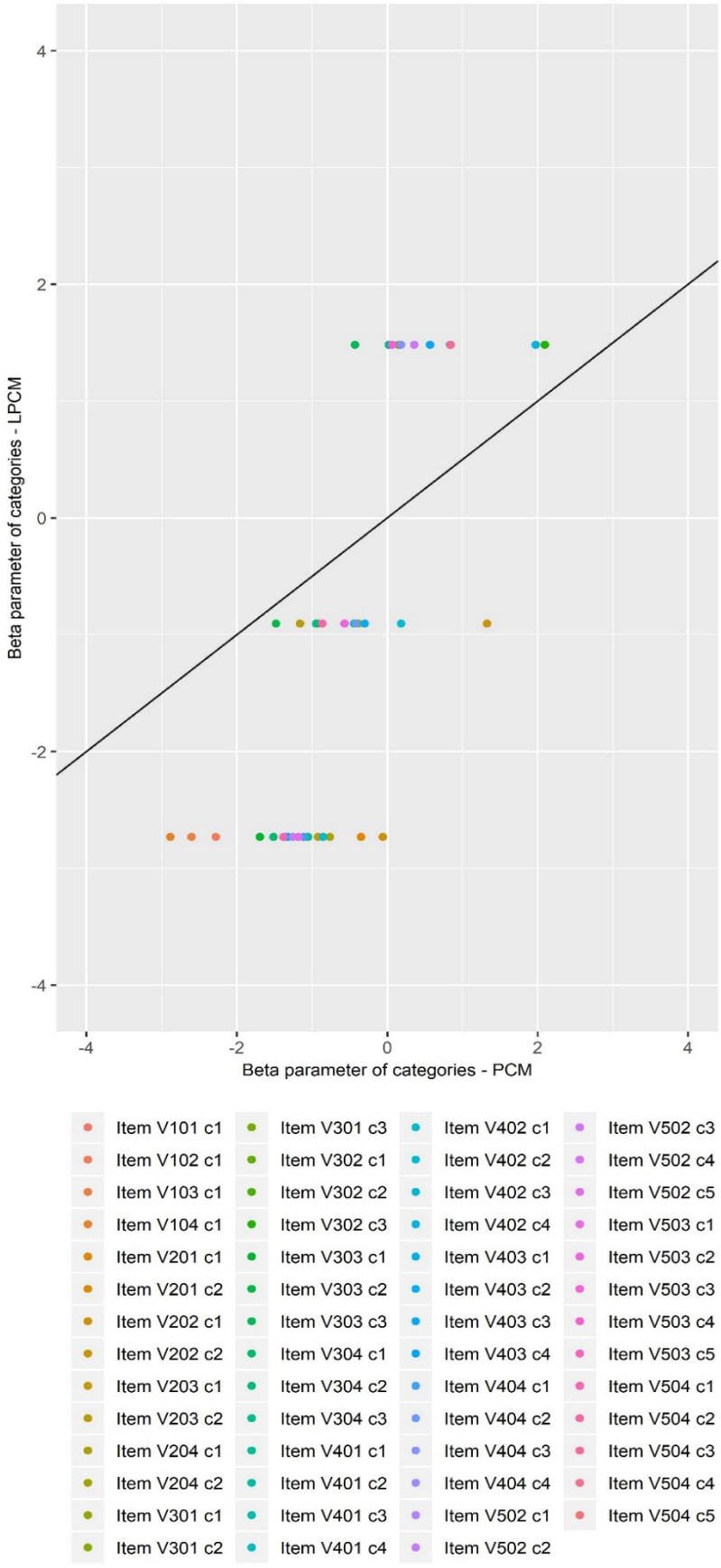


Figure 44. PCM and LPCM beta parameter of categories of WM-V.

Appendix

In a simulation, random q-matrices with different ratios of 0's and 1's were generated and the LPCM calculated with those. The item difficulty parameter of the PCM and the newly calculated LPCM were correlated and the minimal correlation, the median, the mean, the 95th percentile and the maximum correlation determined as can be seen in Table 115. Missing values could not be calculated due to the properties of the artificially generated design matrix.

Table 115

Descriptive statistics for the correlations obtained from simulated weight matrices – WM-V

$\%_1$	Min	Median	Mean	95%	Max
20	-.2408	-.0127	-.0100	.2010	.3469
45	-.2911	.0237	.0230	.2551	.3705

Note: $\%_1$ = occupancy with 1's in the weight matrix; Min = minimum correlation, 95% = 95th percentile; Max = maximum correlation.

In a second simulation, q-matrices were permutated. The results of the descriptive statistics can be seen in Table 116.

Table 116

Descriptive statistics for the correlations obtained from permutated simulated weight matrices – WM-V

Min	Median	Mean	95%	Max
.4620	.7998	.7280	.8817	.8838

Note: Min = minimum correlation, 95% = 95th percentile; Max = maximum correlation.

Table 117

Item difficulty parameter for the PCM and the LPCM – WM-V

Item	Category	Item difficulty parameter PCM	Item difficulty parameter LPCM
V101	1	-1.111	-2.731
V201	1	-2.280	-2.731
V301	1	-2.603	-2.731
V401	1	-2.884	-2.731
V101	1	-0.348	-2.731
V102	2	-0.414	-0.905
V201	1	-0.060	-2.731
V202	2	1.326	-0.905
V301	1	-0.921	-2.731
V302	2	-1.159	-0.905
V401	1	-0.764	-2.731
V402	2	-0.383	-0.905
V101	1	-1.105	-2.731
V102	2	-0.867	-0.905
V103	3	0.152	1.483
V201	1	-1.339	-2.731
V202	2	-0.928	-0.905
V203	3	2.095	1.483
V301	1	-1.694	-2.731
V302	2	-1.478	-0.905
V303	3	-0.427	1.483
V401	1	-1.515	-2.731
V402	2	-0.948	-0.905
V403	3	0.019	1.483
V101	1	-1.052	-2.731
V102	2	-0.439	-0.905

continued

continued

Item	Category	Item difficulty parameter PCM	Item difficulty parameter LPCM
V103	3	0.836	1.483
V104	4	3.423	4.460
V201	1	-0.852	-2.731
V202	2	0.185	-0.905
V203	3	1.975	1.483
V204	4	4.041	4.460
V301	1	-1.318	-2.731
V302	2	-0.298	-0.905
V303	3	0.570	1.483
V304	4	3.254	4.460
V401	1	-1.115	-2.731
V402	2	-0.416	-0.905
V403	3	0.184	1.483
V404	4	2.199	4.460
V201	1	-1.256	-2.731
V202	2	-0.570	-0.905
V203	3	0.360	1.483
V204	4	1.746	4.460
V205	5	3.634	5.977
V301	1	-1.184	-2.731
V302	2	-0.566	-0.905
V303	3	0.065	1.483
V304	4	1.241	4.460
V305	5	2.086	5.977
V401	1	-1.379	-2.731
V402	2	-0.864	-0.905
V403	3	0.841	1.483
V404	4	2.251	4.460
V405	5	2.054	5.977